University of Pavia

Department of Economics and Management

MSc Finance

PREDICTING AND EXPLAINING DOWNSIDE RISK: EVIDENCE FROM S&P 500

Master's Thesis

Maria Lucia Matagliati

Student Number: 554356

Supervisor:
Prof. Giudici Paolo Stefano

July 2026

Contents

Abstract

Traditional risk measures in finance focus mainly on the overall volatility, while investors care more about adverse outcomes, severe losses, drawdowns and crashes. Downside risk is multi-dimensional but often analysed using only one measure at a time.

This thesis predicts firm level downside risk using firm characteristics and assesses whether machine learning can identify important risk drivers. Lastly, predictions are translated into a practical firm ranking framework.

The analysis is done on 415 S&P 500 firms. The predictor variables are measured in 2023 while downside risk outcomes are measured in 2024. This makes the thesis cross-sectional.

Three downside risk dimensions are used (maximum drawdown, downside volatility and the crash indicator) and two machine learning models (LASSO and Random Forest). The results are evaluated mainly through Ranked Gain Accuracy (RGA) and tested for robustness through a subset removal analysis. Additionally, an ablation analysis is done for mechanistic interpretability.

The main findings include that downside risk is meaningfully predictable when using firm characteristics and that machine learning models are able to successfully capture firm level risk patterns. Furthermore, evaluations show that Random Forest performs best for maximum drawdown and the crash indicator, while LASSO is best for downside volatility. Volatility is consistently the most important predictor and financial and market-based variables dominate the forecasting performance. Secondary to financial variables are ESG variables, since they only contribute marginally.

The best predictions from the three downside risk models are converted into percentile scores. These are then normalized and aggregated into multi-dimensional downside risk score where firms are ranked according to the overall downside exposure. The results are robust since the mean and median ranking approaches produce highly similar results: 9 out of the top 10 firms overlap across ranking methods.

The thesis demonstrates that downside risk is multi-dimensional, predictable and can be transformed into a practical ranking framework for investment screening, portfolio monitoring and risk management.

1. Introduction

1.1 Motivation and Background

Financial risk is traditionally measured through metrics that are based on volatility, while investors are often more concerned with downside outcomes than average fluctuations. Additionally, behavioural finance suggests that investors tend to perceive severe losses much more than upside gains, even if they have the same magnitude. This is why events like financial crisis and market crashes highlight the importance of a downside risk. The understanding of the latter is crucial for portfolio and risk managers, as well as for asset allocation.

Lately there has been an increasing use of machine learning models in the financial field due to their ability to capture linear as well as nonlinear relationships and complex interactions. Overall predictive performance and interpretability have to be balanced.

Additionally, there has been a growing importance of ESG considerations in financial markets and their integration into investment decisions. Institutional investors have started placing greater emphasis on sustainability related information and the debate regarding the importance of ESG characteristics has increased.

1.2 Research Problem

The existing literature primarily focuses on traditional risk measures and downside risk has received less attention comparatively. Most studies focus only on one downside risk measure at a time and because of that are not able to capture different aspects of firm vulnerability.

Furthermore, there is mixed evidence regarding relationship between ESG and risk and there is no unified consensus regarding the predictive value of ESG information. On top of this, it is unclear which ESG dimension matters more.

Moreover, many studies focus only on the accuracy of predictions and don't put emphasis on the interpretability of the model. This limits the understanding of which variables are more important in driving downside risk predictions. Practical decision-making applications are often absent as well.

This gap in literature shows the need for an integrated framework, where multiple downside risk measures are predicted through different firm characteristics. The results need to be interpreted with an economic lens and a framework for the practical decision-making has to be created.

1.3 Research Objectives and Research Questions

The main objective of this thesis is to investigate the predictability of downside risk using firm characteristics, including the typical financial variables, measures related to the size of a company, sectors and ESG variables.

The performance evaluates the predictive performance of machine learning methods and their capability in identifying economically significant predictors. Lastly, a practical downside risk decision framework is developed.

The research question for this analysis is “Can downside risks can be predicted using firm characteristics and can these predictions be translated into a practical decision-making framework?”.

Sub-questions that are answered in this thesis are which machine learning model performs best, which variables drive downside risk predictions and whether predictions can be transformed into a practical risk ranking framework.

1.4 Contribution of the Thesis

The first contribution of this thesis is the simultaneous analysis of three downside risk measures, namely the maximum drawdown, downside volatility and the crash indicator. This aspect gives the analysis a multi-dimensional perspective on downside risk.

The second contribution is that two main machine learning models are compared throughout the analysis: LASSO and Random Forest. Through this, their robustness is evaluated as well as their predictive performance.

The third contribution is the inclusion of a mechanistic interpretability framework. An ablation-based analysis is undertaken to identify the economically important predictors. This creates a bridge between prediction and interpretation.

Lastly, a downside risk ranking framework is developed through percentile normalization and ensemble aggregation. In this way, a practical decision-making tool is created that investors can use and adapt based on their preferences and needs.

2. Literature Review

2.1 Downside Risk in Finance

2.1.1 Why Downside Risk Matters

The traditional variance calculations treat the upside and downside fluctuations the same way, even though investors experience these movements differently. Like Markowitz[1] analysed, investors are more concerned when returns are below the target level, meaning that a potential decrease in investment affects them more than a potential increase. The traditional variance treats deviations from the mean equally, independent of the fact if they're positive or negative. This is why Markowitz proposed the semi variance concept, where only observations that are below the mean are observed. Through this, way good outcomes are ignored and only bad outcomes are considered. This argument is basically the antecedent of downside volatility.

Ang, Chen & Xing (2006)[2] investigated further that investors care more about losses than gains. Stocks that perform poorly during recessions and market downturns require a compensation to be held and thus they offer a risk premium, which is found out to be approximately 6% per year. For this reason, downside risk has to be priced and evaluated separately than other financial risk measures, like for example the standard market beta. They explain that risk is not fully captured by the variance or the standard beta. Consequently, the general overall volatility may lose importance for investors confronted to the downside at state volatility. From their paper it is understandable that risk should not be measured only through the overall volatility, but that future risks should focus specifically on bad outcomes. This motivates the analysis of downside variables for the evaluation of future risk, especially for financial investors that care more about losses than gains.

This paper gives this thesis the theoretical foundation for treating downside risk as economically meaningful. It supports the idea, that the risk analysis in negative market states requires a separate evaluation. While the paper applies this idea two downside beta, this thesis extends it to firm level downside outcomes. Additionally, this paper supports the use of financial controls such as beta, size, volatility and momentum.

2.1.2 Different Dimensions of Downside Risk

Downside risk is not captured by one single variable, different measures exist. Some focus on repeated negative fluctuations, some on the worst cumulative loss and some on rare crash events. Thus, different risk measure captures different investor concerns. For this reason, this thesis uses 3 dependent variables instead of only one.

One of the important measures for downside risk is maximum drawdown. It captures an extra dimension and for this reason is often used by investors. It is a common metric for investment decisions and it is frequently used by funds for their allocations, since it is relevant for real world financing. This is what Van Hemert et al. (2020)[3] explained in their paper. It is a measure that

[1] Markowitz, H. M. (2008). *Portfolio selection: efficient diversification of investments*. Yale university press.
[2] Ang, A., Chen, J., & Xing, Y. (2006). Downside risk. *The review of financial studies*, *19*(4), 1191-1239.
[3] Van Hemert, O., Ganz, M., Harvey, C. R., Rattray, S., Sanchez Martin, E., & Yawitch, D. (2020). Drawdowns. *Available at SSRN 3583864*.

analyses the peak-to-trough loss and it captures the worst experience that an investor would occur in that period. It is a path dependent variable and different from volatility, semi variance and skewness in the sense that it measures an investor's worse loss. For example, even if two assets have a similar volatility, they can have very different drawdowns. It is to be noted though, that the evaluation time frame strongly effects the drawdown risk. Because of this, longer evaluation windows naturally increase the probability of observing larger drawdowns. Thereafter, drawdowns capture downside outcomes that are dependent on their path and it represents the losses that investors occur in a specific time frame.

In this thesis, maximum drawdown is included as one of the three downside risk measures since it reflects the investor's worst realised loss path. It is economically meaningful because investors often care about the maximum loss they could incur, not only about fluctuations.

Another measure of downside risk is the downside volatility, which is particularly analysed by Bollerslev, T., Li, S. Z., & Zhao, B. (2020)[4]. They explain that volatility should not be treated as a homogeneous concept, since the return volatility can be composed into good volatility, which represents the positive price movements, and bad volatility, which stands for the negative price deviations. Additionally, good and bad volatility contain different economic information. Furthermore, the authors argument that upside and downside volatility should not be interpreted in the same way and that negative return variations are more relevant for risk assessment. Bollerslev et al. find that the predictive effect of downside volatility is significant after controlling for firm characteristics, suggesting that the measure contains distinct predictive information.

For this analysis downside volatility is included to capture the volatility of negative returns only. The measure differs from maximum drawdown since it captures repeated negative fluctuations and not the worst cumulative loss. For example, a firm may not have one extreme drawdown, but it may still show frequent negative movements.

The returns of stock markets are asymmetric and often large market movements are more declines than increases. This introduces a third measure for downside volatility, which is the risk of a firm to crash. The crash risk is linked to negative skewness, which means that extreme negative returns are more likely than extreme positive returns. This is analysed by Chen, J., Hong, H., & Stein, J. C. (2001)[5] When they forecast conditional skewness in individual stock returns. Their findings include that future negative skewness is strongly related to stocks that have a high trading volume relative to a trend and have strong past returns. Past high returns can reflect bubble like dynamics, since when bubbles build up, the later correction can produce large negative returns. The authors justify that crash like risk can be predicted using previous firm characteristics, which supports the decision making in this analysis.

In this thesis, a crash indicator is included to capture a separate dimension of downside outcomes, namely extreme events. It is differentiated from drawdown and downside volatility because it focuses on tail events. Additionally, for this analysis, it is used as a binary measure.

Further researchers analysed downside risk in adverse economic scenarios. Farago, A., & Tédongap, R. (2018)[6] explored that upside uncertainty and downside risk are not symmetric,

[4] Bollerslev, T., Li, S. Z., & Zhao, B. (2020). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis*, *55*(3), 751-781.
[5] Chen, J., Hong, H., & Stein, J. C. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of financial Economics*, *61*(3), 345-381.
[6] Farago, A., & Tédongap, R. (2018). Downside risks and the cross-section of asset returns. *Journal of Financial Economics*, *129*(1), 69-86.

investors place a greater weight on adverse states. Their paper Argues that downside risk is not represented by only one factor and their model includes several downside related components, creating a five-factor model. They show that their five-factor model performs better than more restrictive models and they come to the conclusion, that volatility is especially important in analysing downside risk, because adverse states can be caused not only by falling returns, but also by rising volatility.

Since downside risk is multi-dimensional, it is suggested to rely on more than one single measure to create a complete assessment of firm vulnerability. This is why three distinct measures are used to predict downside risk and why it is useful to combine multiple downside risk predictions into one ensemble ranking model.

2.2 Predictability of Downside Risk

Downside risk is not purely random since financial markets contain persistent risk characteristics. For example, past volatility, firm level information and market exposure can provide useful data about the future risk a company is going to face. Thus, firm characteristics are able to predict future downside outcomes.

Volatility is one of the most important time varying characteristics of return distributions, which is analysed by Andersen et al. (2003)[7]. It is a key metric used for asset pricing and allocation, as well as portfolio and risk management. The authors argue that volatility can be forecasted using past realised volatility, since it has a strong persistence. Additionally, its predictions improve when high quality information is used.

Their paper justifies the idea that past volatility contains information about future volatility and risk, which supports the use of historical volatility when predicting maximum drawdown, downside volatility and the crash indicator in this thesis. Furthermore, Andersen et al. help explain why volatility emerges as an important variable in the mechanistic interpretability section and why it can be reasonable to remove it when it strongly affects model performance.

Volatility can be studied at different levels and Campbell et al. (2001)[8] study it at 3 specific ones: the market level volatility, the industry level volatility and the firm level volatility. They are able to show that individual stock volatility is not only driven by the overall market, but that firm specific volatility is an important part of the stock risk. Their analysis shows that firm level risk is relevant and that volatility can help to predict future economic activity.

In this thesis, predictions are based on a firm level and not on a market level. The last paper supports the cross-sectional design that different firms have different risk profiles, that firm level volatility is heterogeneous and that firm specific risk cannot be fully explained by just market movements. It incentivizes the usage of firm level variables, such as volatility, beta, momentum and company size.

Firm specific characteristics contain information beyond the aggregate market conditions. Companies have different financial structures, market exposures an operating environment,

[7] Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579-625.

[8] Campbell, J. Y., Lettau, M., Malkiel, B. G., & Xu, Y. (2001). Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *The journal of finance*, *56*(1), 1-43.

which creates a variation in future risk profiles. Therefore, past firm characteristics can be used ask predictors for future downside risk outcomes.

In this thesis, downside risk predictions are treated as a forecasting problem That is solved with machine learning models. Gu, S., Kelly, B., & Xiu, D. (2020)[9] compare machine learning methods in empirical asset pricing, since machine learning can handle large sets of predictors. The authors explain that traditional regressions struggle when there are many predictors and when these are highly correlated, as well as when there are nonlinear relationships and interactions between the predictors. Machine learning models are able to solve these problems by reducing overfitting through regularisation and model selection. Additionally, they demonstrate that tree-based models and neural networks perform especially well. However, they also note an important limitation: machine learning predictions are solely measurements, they do not automatic explain economic mechanisms.

This paper sets the modelling foundation for this thesis, since machine learning is used for forecasting rather than traditional regressions. Gu et al. support the use of LASSO and Random Forest, since LASSO fits the idea of regularisation and variable selection, while Random Forest analyses nonlinear relationships and interactions. Their paper also supports the inclusion of many predictors at the same time.

Overall, the literature supports the predictability of risk related outcomes. The authors mentioned above show that volatility is persistent and forecastable, that a firm level approach is economically important and that machine learning can extract predictive signals. Together, these papers justify the empirical setup for this thesis.

2.3 Literature Gap and Positioning of the Thesis

From the past sections it is clear that the literature already explained why downside risk is economically important and multi-dimensional. Additionally, the predictive ability of volatility and firm level characteristics has already been established, as well as the usefulness of machine learning in financial prediction.

Nevertheless, most studies focus on a single downside risk measure and there is limited evidence that combines the following three downside measures: maximum drawdown, downside volatility and the crash risk. Furthermore, there is limited evidence on ESG variables as predictors of downside risk. The machine learning literature often prioritises prediction accuracy instead of interpretability and there is incomplete data on which variables drive downside risk predictions. Moreover, a practical decision-making framework is rarely developed.

This thesis serves to predict three downside risk dimensions simultaneously, combining multiple firm characteristics. The result and compared through different machine learning approaches and the economic drivers are investigated through mechanistic interpretability. Lastly, a multi-dimensional downside risk ranking framework is developed to translate predictions into an investor-oriented decision tool.

[9] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.

3. Data and Variables

3.1 Dataset Description

The dataset for this analysis consists of the S&P 500 companies in the year 2023. The S&P 500 is a market-weighted index of leading publicly traded U.S. companies and it is widely used as a benchmark for the performance of the U.S. equity market.

An excel worksheet from Kaggle was used for the ESG values of S&P 500 companies for the year 2023, as well as for the market capitalization and number of employees of companies. Even though the index includes 500 companies, for this analysis the clean final sample consists of 415 firms from the Index.

The 85 missing companies were discarded in the evaluation because of missing values or their removal from the index in the following year (2024). Mainly, companies were not included because of absent ESG scores. Additionally, a minimum number of observations was required for the computation of returns and risk measures.

The 415 S&P 500 companies were analysed cross sectionally for the financial year 2023, seeing how their firm characteristics and risk measurements influenced the year 2024.

3.2 Feature Construction

The analysis contains a total of 20 variables, including continuous variables, ESG scores and sector dummies.

The features can be grouped into 4 categories: risk/ financial variables, firm characteristics, ESG variables and sector dummies.

The same feature matrix was used across all models to ensure comparability of results.

The risk variables include the volatility in 2023 (Volatility_2023), the market beta (Market_Beta) and the momentum (Momentum_2023).

The momentum was computed using 2023 adjusted daily closing prices from Yahoo Finance with the following formula:

$$Momentum\ for\ stock\ i = \frac{Price\ (end)}{Price\ (start)} - 1$$

The momentum represents the firm's total stock return in the whole year of 2023, meaning that if the momentum was positive, the stock increased in 2023, while if the momentum was negative, the stock declined during the year. The metric was calculated using adjusted closing prices for 2023, since they account for dividends, stock splits and corporate actions, reflecting the true total return one would have holding the stock. If raw prices would have been used, dividends would be ignored, returns would be underestimated and the comparisons across firms would be biased. Through the usage of adjusting closing prices, total shareholder return for 2023 is computed.

Volatility is the standard deviation of the daily returns in 2023 and is a standard proxy for total risk. It measures how much returns deviate from their average. The daily returns of each company are

computed as the relative change in adjusted closing prices between two consecutive trading days.

$$Daily\ returns\ of\ a\ stock = \frac{Price_t - Price_{t-1}}{Price_{t-1}}$$

Volatility is then defined as defined as:

$$Volatilty\ of\ stock = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(daily\ return - average\ return)^2}$$

The volatility uses adjusted prices and measures how much the stock price fluctuate. A high volatility means that a company has large swings and fluctuations, signalling its riskiness, while a low volatility company is more stable and safer to invest in.

Volatility is a key determinant of downside risk, since stocks with a higher return dispersion are more likely to have bigger negative movements.

The market beta was calculated with the covariance between the daily returns of a stock and the daily return of the market (rm) and the variance of the daily returns of the market. For the daily returns of the market, SPY was used. SPY is an ETF tracking the S&P 500 index and is used as a benchmark for the U.S. equity markets. Therefore, the market return is the return of SPY. The formula is:

$$ß = \frac{Cov\ (ri, rm)}{Var\ (rm)}$$

The market beta measures how much a stock moves with the market. For this analysis, the standard CAPM beta formula was used to capture systematic risk. A higher market beta signals that a stock is more exposed to market crashes.

The firm characteristic variables are the market capitalization (log_marketcap) and the number of employees in the companies (log_employees).

The market capitalization variable is used to measure the firm size. The following formula was used to compute is:

$$\text{log\_marketcap}\ of\ firm = \ln\ (1 + MarketCap\ of\ the\ firm)$$

The number of employees of a company is used as another proxy for firm size as a way to add another dimension instead of only having the financial size through market capitalization. The formula used is:

$$\text{log\_employees}\ of\ firm = \ln\ (1 + Employees\ of\ the\ firm)$$

The raw variables are taken for both measures from the clean Kaggle dataset, but then log transformed through Python. This decision was taken since both the market capitalization and the number of employees are skewed. Through a log transformation, a reduction in skewness was achieved in addition to stabilizing the variance and improving the correct model performance. The distributions were made more normal.

The ESG variables consist of the total Environmental-Social-Governance Score (ESG_Score) and the 3 single scores, environmental (Env_Score), social (Soc_Score) and governance score (Gov_Score). In addition, an ESG risk score variable was used as well (ESG_Risk_Score), which

stands for exposure to ESG risks that could financially affect the firm. The higher the ESG risk score, the higher the possible ESG related risk.

These variables are all taken from the Kaggle dataset for the year 2023. They measure the environmental performance, the social responsibility, the governance quality and the overall ESG exposure.

ESG variables are included in the analysis since ESG is increasingly studied in finance and may affect risk, stability and downside exposure. Through the inclusion of non-financial variables it gets tested whether ESG can add predictive power beyond classic financial measurements.

Lastly, sector variables are encoded as dummy variables and included as industry controls. The Global Industry Classification Standard (GICS) is used to classify the 415 S&P companies in 11 different sectors: Financials, Health Care, Industrials, Information Technology, Consumer Discretionary, Consumer Staples, Energy, Utilities, Materials, Communication Services and Real Estate. For the analysis, the Real Estate sector was omitted and used as a reference category against which all other sectors were compared to. This avoids perfect multicollinearity between dummy variables, as the sum of all dummies would otherwise equal one.

Sector dummies are used to indicate which industry the firm belongs to. If the firm belongs to a sector it is denoted with 1, otherwise 0.

$$Dummy = \begin{cases} 1 & if\ firm\ is\ in\ sector \\ 0 & otherwise \end{cases}$$

This transformation is done to transform categorical variables into numerical form for models. They are used to control for industry differences. Without sector dummies, the model could be biased since sector effects could be attributed to other variables in the model. Different sectors have different risk, for example the tech sector is known to be more volatile than the utilities sector. Without sector dummies, the model could have predicted high volatility from firm characteristics, while it actually comes from a sector. Through the inclusion of sector dummies as control variables, industry specific risk differences can be accounted for.

3.3 Outcome Variables

As outcome and dependent variables for this analysis, 3 measurements are used to analyse 2024 downside outcomes. These 3 metrics are the maximum drawdown (MaxDrawdown_2024), the downside volatility (DownsideVol_2024) and the crash indicator (Crash_2024). These 3 together allow the downside risk to be evaluated multidimensionally, capturing different dimensions.

The predictors are based mainly on information from 2023 while the outcomes are measured in 2024, creating a forward-looking analysis.

The downside outcomes don't rely on a single prediction framework. The maximum drawdown and downside volatility are continuous variables, as they can take a wide range of numerical values, and are modelled through regression techniques. The crash indicator is a discrete binary variable that is analysed through a classification problem. This way the evaluation predicts on one side through the regressions the magnitude of risk, on the other side it looks at extreme events through the classification.

Through the usage of multiple downside risk measures, the result is more comprehensive and complete, capturing different aspects of risk. Maximum drawdown reflects cumulative losses,

downside volatility captures the variability of negative returns, and the crash indicator identifies extreme tail events.

3.3.1 Maximum Drawdown 2024

Maximum drawdown measures the worst peak-to-trough loss that a stock experienced during the year 2024. It is defined as the largest decline from a peak to a following low within the year, accounting for time ordering of the prices. It represents the worst loss that an investor would have incurred if he had bought the stock at a high and sold at the following low.

The code tracks the stock price day by day, keeping the highest price observed up to date. Each day, the drawdown is:

$$Drawdown_t = \frac{Price\ at\ t}{\max(Price_1,\ \ldots,\ Price_t)} - 1$$

The adjusted closing prices of a firm each day are used. The maximum drawdown is then the minimum value of the drawdown series:

$$MaxDrawdown\ for\ a\ firm = \min_t(Drawdown_t)$$

Drawdowns are losses, which means that the values are negative. For this reason, the worst loss corresponds to the most negative value. The larger the negative value is, the larger the loss from a previous peak is and the higher the downside risk is.

Drawdown was only computed for observations that have at least 50 valid price values, otherwise the drawdown variable would be based on too few days and would result unreliable.

Maximum Drawdown captures investors fears by measuring large cumulative losses. While volatility measures fluctuations, drawdown measures the real loss path, because a stock can be volatile but recover quickly from negative periods. Maximum drawdown shows the worst realised decline over the year, it focuses on capital loss.

3.3.2 Downside Volatility 2024

Downside volatility measures how much a stock fluctuates on “bad days only”, meaning on days when the returns are negative. The difference to normal volatility is that general volatility uses all returns, negative and positive ones, while the downside volatility used for the analysis focuses only on negative returns.

Downside volatility was computed with daily returns from adjusted prices, analogously to the volatility variable (see section 3.2), but keeping only daily negative returns. The downside volatility is then computed as the standard deviation of the negative daily returns.

The code is set up to require at least 50 valid return observation and at least 10 negative return observation, otherwise downside volatility cannot be correctly estimated if there are not enough negative days.

Downside volatility is used as a more focused risk measure than normal volatility. Investors usually care more about downside movements than positive surpluses, reason for why downside volatility was chosen as a variable instead of standard volatility, since it captures the instability of losses instead of the general price movement.

3.3.3 Crash Indicator 2024

The crash indicator identifies firm that had extremely low annual returns in 2024. It shows the 10% of the worst performing companies in that year, separating normal firms from extreme underperformers.

The crash indicator is computed through the annual return for each company using adjusted closing prices:

$$Return\ of\ a\ stock_{2024} = \frac{Price_{end\ 2024}}{Price_{start\ 2024}} - 1$$

Then the crash indicator was calculated with the 10th percentile of the annual return distribution:

$$q_{0.10} = 10th\ percentile\ of\ Return_{2024}$$

The threshold was approximately -18.5%, meaning that firms with an annual return below or equal to the cut-off point were classified as crash firms. For this metric as well, the code was set up to require at least 50 valid price observations to be able to compute a correct variable.

The crash indicator is a binary variable, resulting in a company being assigned 1 if it is a crash firm, while it being assigned 0 if it is a non-crash firm.

$$Crash\ firm_{2024} = \begin{cases} 1 & if\ Return_{2024} \leq q_{0.10} \\ 0 & otherwise \end{cases}$$

The code classified 42 firms as crash firms, signalling that they represent the 10% of the sample with the lowest return in 2024.

Through this binary outcome, the crash indicator creates a tail-risk classification problem, because instead of predicting the size of loss of a company, the model predicts whether a firm belongs to the worst tail of the return distribution. This is useful in finance since investors and risk managers often prefer to identify risky and underperforming firms rather than knowing the exact return.

3.4 Descriptive Analysis

3.4.1 Cross-Sectional Perspective

The data set consists of the S&P 500 firms and there is one observation per firm available. The explanatory variables are measured in the 2023 outcome period, while the downside risk measures are calculated for 2024. The comparison is it conducted across firms rather than across times and the focus is on cross-sectional heterogeneity. The objective is to identify the differences in downside risk between the companies, with the help of firm characteristics and machine learning prediction models.

3.4.2 Summary Statistics

Outcome	Mean	Median	Std. Dev.
MaxDrawdown_2024	-0.2053	-0.1792	0.1038
DownsideVol_2024	0.0118	0.0108	0.0050
Return_2024	0.1406	0.1258	0.2812

From the table it can be seen that the average maximum drawdown is around 20%, whereas the median is less severe. This indicates that a subset of firms experienced particularly large losses. There is also a considerable difference in drawdown outcomes.

On the other hand, downside volatility is concentrated around relatively low levels, where the mean is slightly above the median. This indicates a mild right skewness.

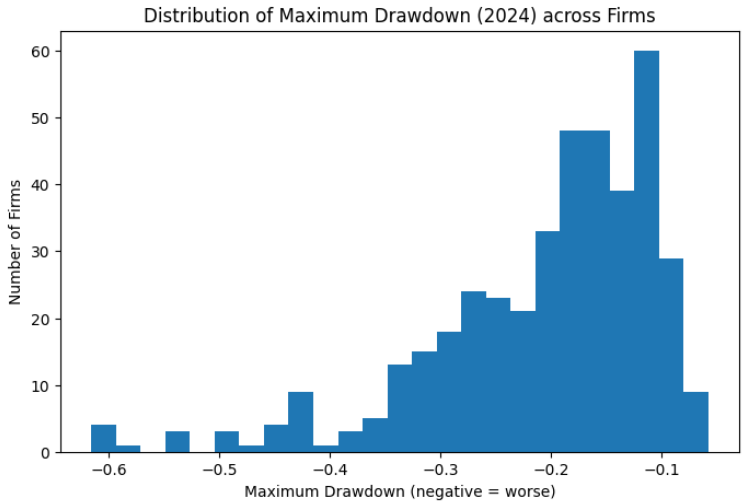
Furthermore, the annual returns show large dispersions among outcome variables, indicating that there are important differences between the performance of companies during 2024. It is observable though, that both the mean and the median returns are positive, which shows that most firms generate a positive return.

In addition, 10% of the firms were classified as crash firms for the crash indicator. This is consistent with the bottom-decile crash definition employed in this thesis.

This summary indicates that the sample is not homogeneous and that there are meaningful differences between the firms. Moreover, there is sufficient dispersion for predictive modelling. The downside risk measures don't concentrate only around one single value and in the sample, there are firms present with both low and high downside exposure. This supports the use of a cross-sectional prediction framework and it indicates, that ranking firms according to the downside risk is economically meaningful.

3.4.3 Distribution of Downside-Risk Measures

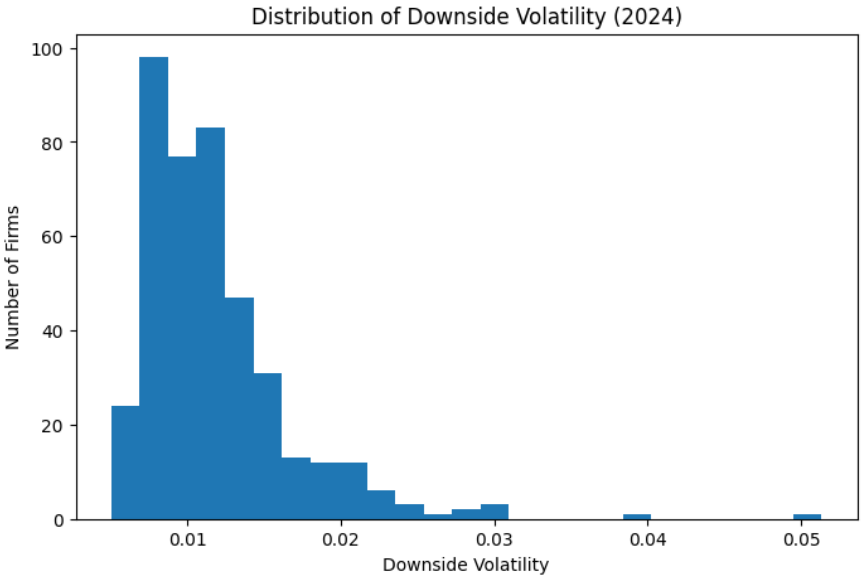
Complementing summary statistics, the following section provides visual examinations of downside risk measures. It focuses on skewness and dispersion, identifying extreme observations.



The graph above represents the distribution of maximum drawdown. A concentration around moderate drawdown values is observable, where the majority of firms cluster. This shows that most firms experienced losses of moderate magnitude.

From the summary statistics table it is known that the mean is more negative than the median, which shows evidence of left skewness. There are firms that present very severe drawdowns. Overall, the losses are not evenly distributed.

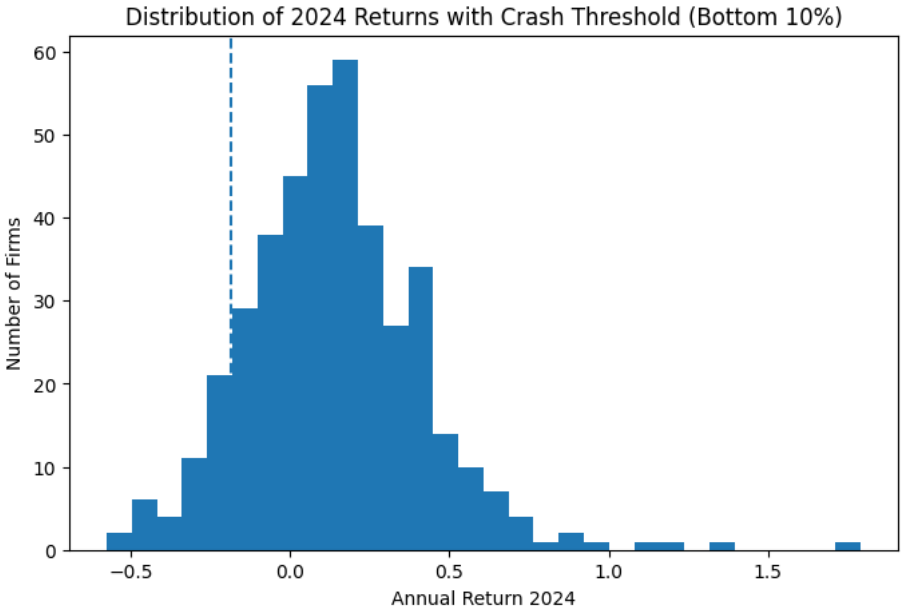
Economically, this shows that some firms are more exposed to bad market conditions and that severe drawdowns are concentrated Between a subset of firms.



The downside volatility distribution presents a high concentration around relatively low volatility levels. This shows that the majority of firms exhibit a moderate downside volatility.

Here the mean only slightly exceeds the median, which results in a mild ride skewness. Only a limited number of firms usually display high downside volatility. As it can be seen in the graph, a smaller group of outlier firms contributes to the upper tail.

Economically it can be said that most firms experience a relatively stable downside volatility, but that a smaller subset of firms experiences higher fluctuations.



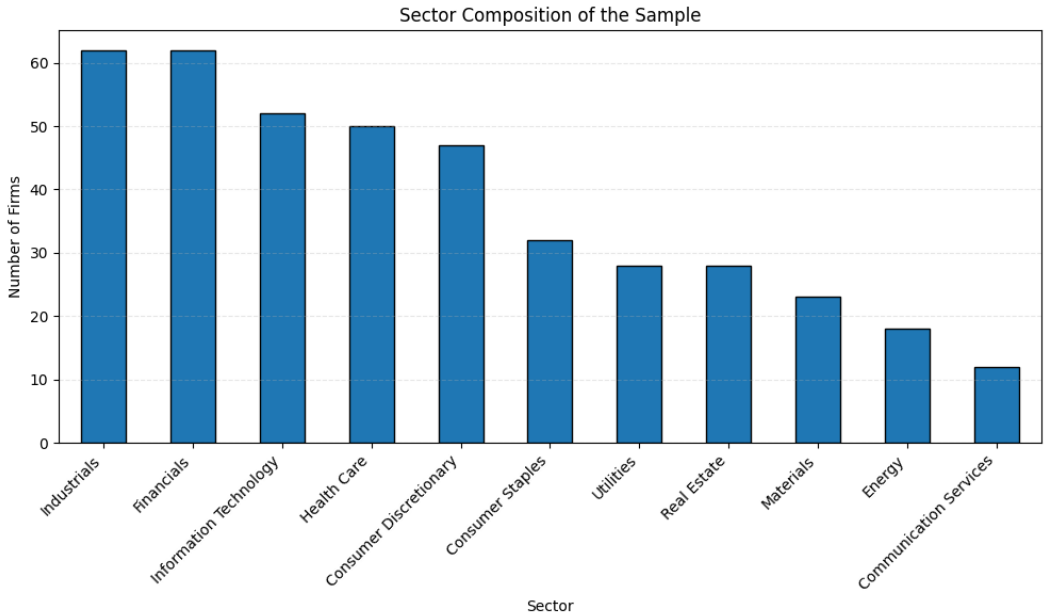
The distribution of annual return shows that there is a broad distribution of firm performance and that the average returns are positive. Substantial variations are present though across firms.

The mean exceeds the median, which is an indication of positive skewness. It shows that there are companies with an exceptionally strong performance. The dispersion is the largest spread between the outcome variables, which shows a strong cross-sectional heterogeneity.

The vertical threshold line separates the crash firms from the non-crash firms. It represents the bottom decile return cut off.

Economically, firm performance differs substantially, even within the same market environment. There exists both strong winners and severe underperformers, which justifies the use of a downside risk prediction framework.

3.4.4 Sector Composition



The S&P 500 index is composed of firms from multiple industries. This sector distribution provides an overview of the sample composition. It shows that firms operate in different economic environments. As a consequence, sector composition is important to understand the cross-sectional heterogeneity of the sample since different sectors have different business models and are exposed to different risks.

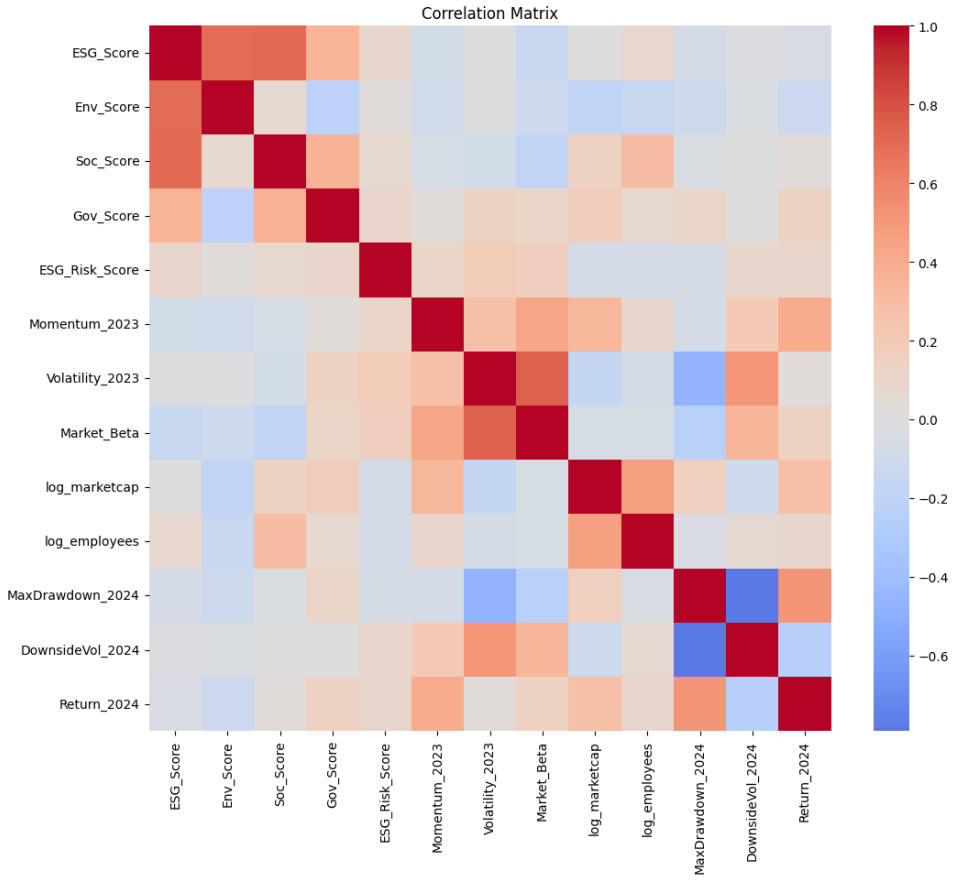
From the graph it is understandable that the sector representation is not evenly distributed. Some sectors account for large proportion of the observations, while others account only for a minimal percentage. For example, the industrials, financials and information technology sectors are the most represented ones in the sample, while energy and communication services are the least represented ones. The broad sector coverage of the sample though, increases its representativeness.

This graph shows that the sample is not only concentrated in one single industry. Contrarily, multiple economic activities are represented, which improves the generalisability of empirical findings. It allows the comparison across different companies.

Sectors differ substantially in economic characteristics, since they are exposed to different business cycles, competitive environments, regulatory frameworks and risk profiles. For example, financial firms are more exposed to credit and market conditions, while technology firms are more exposed to innovation and valuation risk.

Consequently, the inclusion of sector dummies in this thesis is important, since the sector had resonated you can influence downside risk outcomes. Different firms that operate in different sectors face different sources of risks, and because of this, sector effects may partly explain the variation in downside risk measures.

3.4.5 Correlation Matrix



The correlation matrix provides an overview of the linear relationships. It examines the association between explanatory variables and downside risk measures. It is useful for identifying patterns and to detect potential multicollinearities.

The strongest positive correlation is observed between the total ESG score and the individual ESG pillars. Between these, the strongest relationship happens between the ESG score and the environmental score. They all have a positive relation. This result is expected, since the aggregate ESG score incorporates the information from the individual pillars.

Furthermore, a positive correlation can be observed between the logarithm of the market capitalization in the logarithm of the employees. This is due to the fact that larger firms generally have more employers.

Additionally, another strong positive correlation is present between the volatility and the market beta. This can be explained since firms that have a higher systematic market exposure tend to exhibit higher volatility.

Moreover, there is a positive relationship between the return in 2024 and the maximum drawdown in 2024. It can be explained since firms with stronger returns generally experience less severe drawdowns.

Is important to know that there is a strong negative relationship between the outcome variables maximum drawdown and downside volatility. This result can be due to the fact that firms which present a more severe drawdown, tend to suffer from a higher downside volatility.

From the high aggregate ESG score, which is very correlated with the ESG subcomponents, the potential multicollinearity issue can come up. As mentioned before, this result is expected since they measure related sustainability dimensions. The multicollinearity problem is not a major issue for this analysis, since LASSO Includes regularisation, which reduces the influence of redundant predictors. Through the variable selection process, multicollinearity is mitigated. Also, Random Forest is less sensitive to correlated variables. Because of these reasons, the correlation levels don not appear sufficiently extreme to invalidate the analysis.

4. Methodology

4.1 Overview of the Methodological Approach

The objective of this thesis is to predict and rank future downside risk through three complementary downside measures: maximum drawdown, downside volatility and crash indicator. The prediction is based on firm-level variables that can be sorted in four categories: financial risk variables, firm characteristics, ESG variables and sector dummies.

The analysis is conducted in a cross-sectional setting and is carried out with a forward-looking approach. The predictor variables are constructed using 2023 information, while outcome variables capture realized downside risk in 2024.

The three downside risk measures include both continuous and binary outcomes. Maximum drawdown and downside volatility are modelled as continuous outcomes, while the crash indicator is built through a classification problem. Additionally, two machine learning models are employed to estimate these outcomes: LASSO and Random Forest.

For model performance evaluation, Ranked Gain Accuracy (RGA) is used, a ranking-based evaluation metric. RGA is used in machine learning to assess the performance of regressors or classifiers. It focuses on the ordering of predictions, rather than the exact predicted value, as is common in traditional prediction metrics. This approach is particularly relevant in financial contexts, where the ability of identifying high-risk firms is often more important than predicting exact outcome values.

Once computed the baseline RGA, additional analyses are conducted by recomputing the RGA under subset removal, prediction noise and feature removal. These include assessing model performance across different subsets of firms, evaluating the stability of predictions under perturbations and analysing the contribution of individual variables. Through these analyses, the components provide a thorough framework for understanding how well models perform, as well as assess how reliable and interpretable their predictions are.

4.2 Predictive Models

All models are estimated using K-fold cross validation to ensure that the predictive information is evaluated out of sample. For this procedure, the sample is divided in K folds that are mutually exclusive. Then, in each iteration K-1 folds are used to estimate the model, and the remaining fold is used for testing. This process is repeated until every observation has been used once as a test. For this thesis a 5-fold cross validation is chosen.

Additionally, a random state is specified in the modelling procedure, namely “42”. This is done since machine learning algorithms contain random elements, like for example the fold allocation in cross validation. Through a fixed seed, it is ensured that the models have identical train/ test splits and that the results are the same when the code is rerun.

For the crash indicator, the models use a stratified K fold instead of an ordinary K fold, Otherwise some folds may contain very few or too many crashes. Through the stratified K fold, each fold receives approximately the same class proportions as the full sample. In the case of this thesis each crash fold remains close to 10% and each non-crash fold close to 90%.

4.2.1 LASSO

LASSO is a linear regression model that assumes a linear relationship between the predictors and the outcome. The core idea behind it is that the model predicts the outcome as a weighted sum of input variables, where the coefficients represent the marginal effect of each feature.

The LASSO model includes regularization, which is a regression technique used in machine learning to prevent overfitting and improve model interpretability. Regularization is achieved through the insertion of an L1 penalty term on coefficients. Therefore, the objective function balances the fit to the data and the size of coefficients. The L1 penalty shrink coefficient towards zero with some of them being exactly equal to zero.

Through the sparsity feature of the lasso model, n coefficients are set equal to zero. Consequently, the model selects only a subset of variables. This mechanism acts as an automatic feature selection and reduces the dimensionality of the model.

Before the estimation, the variables in the LASSO model get standardised, which means that their mean is set to zero and their variance to one. Standardisation ensures that the penalty treats all variables equally, otherwise variables with a large scale would dominate the model.

Interpretability with LASSO is straightforward, since non 0 coefficients are directly interpretable. The sign of the coefficient must be interpreted relative to the outcome variable. For downside volatility and the crash indicator, if the sign of the coefficient is positive, that variable increases predicted risk, while if the sign of the coefficient is negative, the variable decreases predicted risk. For maximum drawdowns, the opposite holds, since it is expressed as a negative loss measure. The more negative the drawdown outcome is, the worse is the risk.

LASSO is estimated twice with two separate specifications, K=5 and K=10. In the first specification the model is constrained to retain approximately five predictors, while in the second one the model is constrained to retain approximately ten. The purpose of using different sparsity levels is to assess how predictive information changes as the model complexity increases.

The magnitude is also to be analysed, since the larger the coefficient the stronger the influence of that variable on the model. Through the LASSO modelling, the identification of the key drivers of downside risk is straightforward.

The LASSO model has multiple strengths. First, it is a simple and transparent model, which allows a straightforward interpretation. Furthermore, it highlights the most relevant variables, which is a useful characteristic for mechanistic interpretability. Moreover, it is robust in the presence of many predictors, an important quality when the analysis includes numerous variables.

The model's limitations are the assumption of linear relationships. It is not able to capture non-linear effects and interactions between variables. Consequently, it may oversimplify complex financial relationships.

Concluding, LASSO is the baseline model of this thesis, providing an interpretable and clear benchmark for the comparisons with other models. It allows the identification of the main downside risk drivers and complements the Random Forest, the second model used as an opposite approach with more modelling complexity.

4.2.2 Random Forest

Random Forest is the second machine learning model used to evaluate the outcomes. It is an ensemble learning method that combines predictions from multiple distinct decision trees into an ensemble (forest) to improve predictive performance. It constructs a large number of trees, each trained on a subset of the data, instead of relying on a single decision tree. The final result is the average of the predictions across trees for regression, while for the classification outcome the model predicts crash probabilities, which are then used for RGA.

The intuition behind decision trees is that the model splits the tree recursively into data subsets. The splits are chosen to reduce the prediction error. Each path corresponds to a combination of conditions on variables.

Contrary to the LASSO model, Random Forest does not assume a linear relationship of the data. Consequently, it can capture curved relationships and different threshold effects. It includes interaction effects, which refer to the model's ability to identify complex non-linear dependencies between two or more input variables that influence the target outcome together. There is no need to pre-specify interactions, since the model inherently detects these interactions automatically through his hierarchical branch-based decision structure.

Additionally, Random Forest does not have specific variable specification, since the model automatically handles variance important internally. Instead, the complexity of Random Forest is controlled through hyperparameters. The sample is still split into five folds and for each iteration 5 folds are used for training and one is held out for testing. On top of this, to choose optimal hyperparameters, an inner 3-fold cross validation is performed. Each outer training sample has another cross-validation procedure done, where the inner cross validation is split into three folds. These test different candidates for hyperparameters, such as the maximum depth or the number of estimators.

Random Forest is a highly robust model due to the algorithm's ability to maintain high prediction accuracy despite noisy data, outliers and complex datasets. As an ensemble method, it reduced the variance through the aggregation across many trees. This creates a model which is less sensitive to data noises and individual observations.

In addition, it is a highly flexible model, that can approximate functional relationships and adapts to the data structure. It can handle simultaneously non-linearities, interaction and the heterogeneous effects of variables.

The model doesn't require standardisation, since tree-based models depend on splits, not on scales, so the scaling of variables does not affect splits. The raw feature values can be used directly.

Random Forest is considered a "black-box model", since it doesn't include a simple coefficient interpretation. Its interpretation is limited due to difficulties in isolating individual variable effect. A more accurate interpretation requires additional methods.

Concluding, Random Forest is a good benchmark for predictive performance and it is able to capture complex pattern that simpler models, such as LASSO, cannot. Additionally, it is useful to test whether non-linearity improves the models ranking.

4.2.3 Comparison

LASSO being a white-box model, had a more transparent structure. Its coefficients are directly interpretable and there is a clear relationship between inputs and outputs. Random Forest, on the other hand, is a black-box model. Its internal structure is more complex and the predictions result from many trees, which doesn't allow for simple mapping of variable outcomes.

The two models complement each other and represent a trade-off between interpretability and flexibility. LASSO wins on interpretability, since it's an easier to explain model and it can be identified which variables matter more as well as the direction of effects. These characteristics are especially useful for economic interpretations. On the other hand, Random Forest wins on flexibility. The mode captures complex relationships and allows for non-linear effects and interactions. It is better at modelling real-world complexity.

A second trade-off is remarkable between LASSO and Random Forest. Lasso shows a higher bias due to its simpler model structure, but a lower variance, which can result in the model underfitting complex pattern. On the opposite, Random Forest presents a lower bias since its model is more flexible, and a controlled variance, achieved through averaging. It is a better fit for complex data.

Through the setup for this analysis, LASSO relies on a few strong predictors and has a sharp dependence on key variables, while Random Forest distributes the importance across more variables and uses information smoother.

For this thesis, LASSO is useful for identifying the main drivers of downside risk and interpreting the variables correctly, while Random Forest is useful for capturing the full structure of the data and is potentially better at ranking the performance of the variables.

The comparison of both of these models allows the evaluation of simple versus complex models as well as a linear versus non-linear structure. Through these different models it is tested whether additional complexity improves the ranking of variables and it provides robustness since results are not driven by only one model type.

4.3 Ranked Gain Accuracy: RGA

The RGA computation is used as the primary model evaluation metric in this analysis. Its purpose is to obtain a ranking-based performance metric that evaluates whether the predicted risk scores are able to correctly order firms by realized risk. Thus, the focus is on the relative ranking, not predicting the exact value. Furthermore, RGA is especially relevant in finance, since investors often need to identify high-risk companies. The exact risk values are less important, the focus is on knowing which firms are riskier. RGA is used for both continuous outcomes (MaxDrawdown_2024, DownsideVol_2024) and the binary outcome (Crash_2024), allowing for a comparison of all three metrics outputs through ranking.

The RGA formula for the code is taken from Giudici's and Kolesnikov's paper "SAFE AI metrics: An integrated approach" (2025)[10]. The function rga_cramer(y,yhat) is used, which is computed through

$$RGA = 1 - \frac{CvM(y, \hat{y})}{G(y)}$$

[10] Giudici, P., & Kolesnikov, V. (2025). SAFE AI metrics: An integrated approach. *Machine Learning with Applications*, 100821.

$G(y)$ stands for the Gini coefficient of the true outcome. $CvM(y, \hat{y})$ is the weighted Cramér–von Mises distance between Lorenz and concordance curves. y is the realized outcome, while $\hat{y}$ is the predicted outcome. The returns of the function are NaN if Gini is either not finite or equal to zero, or if CvM is not finite. This means that RGA is only meaningful when the true outcome has enough dispersion to rank firms.

The RGA uses two specific inputs. Y is the true realized outcome of downside risk observed in 2024, while yhat is the model prediction generated by LASSO or Radom Forest. Once inserted in the code, both arrays are converted into one-dimensional numerical arrays. In case of missing values, these are removed before the computation. Through this calculation, RGA compares the ranking implied by prediction with the ranking implied by realized outcomes.

The intuition behind the computations is that the model ranks firms well if the companies predicted as risky also have a high realized downside risk. If the model ranks firms poorly, then the predicted ordering does not match the realized risk ordering. RGA rewards models that place actually risky firms toward the high-risk end of the ranking. It does not require the model to perfectly predict the exact numerical value. RGA is a useful metrics for when the economic objective is prioritization, such as classifying riskier firms, deciding which companies should receive more attention or which entities are more exposed to downside risk outcomes.

As mentioned before, the numerator of the RGA equation is the weighted Cramér–von Mises distance between Lorenz and concordance curves. The Lorenz curve is computed from the true outcome y, where the true outcome values are sorted by realized risk. Then, the cumulative share of realized risk is calculated. The curve represents the ideal ordering based on actual outcomes. Summing up, the code sorts y, computes the cumulative sum and then divides by the total sum.

Intuition wise, the Lorenz curve shows how realized downside risk accumulates when firms are ordered by actual outcomes. It represents the concentration of risk in the sample. In this analysis, if the downside risk is concentrated in a few firms, the Lorenz curve reflects the inequality. This is a key part for RGA, because it normalizes performance by the amount of dispersion in the true outcome.

The concordance curve is computed using both the true outcome y and the predicted outcome $\hat{y}$. For the curve, the companies are sorted according to their predicted value. After that, the cumulative realized risk is calculated following the predicted order. The concordance curve shows how much of the realized risk the model is able to capture in his ranking. Summing up, the code sorts observations by yhat, then computes the cumulative sum of y following the order of observations generated in the previous step and then divides by the total sum of y.

The insight is that the Lorenz curve serves to order the realized risk by the actual risk. On the other hand, the concordance curve arranges realized risk by predicted risk. If the ranking of the prediction is good, then the concordance curve is close to the Lorenz curve. Otherwise, the concordance curve is far from the Lorenz curve.

Putting all the passages together, the Cramér–von Mises distance $CvM(y, \hat{y})$ is the weighted distance between the Lorenz curve and the concordance curve. If the model shows a small distance, the predicted ranking is close to the ideal ranking, while if the distance is larger, the predicted ranking is far from the ideal ranking. The weights depend on the contribution of the true outcome to the total outcome. In the code, the absolute distance between the curve is multiplied by weights and summed across observations.

Summing up, the CvM distance measures the ranking error, not the error that is present in predicted values. It is the error in how realized risk accumulates under predicted ordering. A lower CvM indicated a better ranking.

Moving on to the denominator of the RGA, it shows the Gini coefficient of the Lorenz curve. Gini measures the dispersion and inequality of the true outcome. It serves to normalize the CvM distance by the amount of variation in realized risk. If the true outcome has no dispersion, Gini will be equal to zero. This leads to a non-meaningful ranking, which the code returns as a missing value.

It follows that ranking only matters if firms differ in realized risk, since if all companies had the same downside risk, then no meaningful ranking would exist. Through the Gini normalization, the RGA becomes scale-adjusted and allows the comparison across outcomes, even if they have different distributions.

Recapitulating the RGA formula, CvM/G is the normalized ranking error, while $1 - CvM/G$ is the ranking accuracy. The higher the RGA score, the better the ranking of downside risk is. With an RGA close to 1, the model predictions are able to produce a ranking close to the realized risk ordering, while with a RGA close to 0, the ranking quality is weak. In case of a negative RGA, it is interpreted as a ranking worse than the normalized benchmark.

It is important to note, that for maximum drawdown, the outcome variable itself is negative, so the more negative the drawdown value is, the higher its downside risk is. For downside volatility and crash probability, it is the opposite, since larger values indicate a higher risk.

For the two regression outcomes maximum drawdown and downside volatility, the RGA model produces a continuous predicted value. The RGA compares the true continuous outcome y to the predicted continuous risk score $\hat{y}$ and no threshold is used. The model is not judged by mean square error or R^2, but through the ranking quality.

On the other hand, for the classification outcome crash indicator, the outcome is binary: 1 if the company is a crash firm, 0 if the company is a non-crash firm. Here, the model produces the predicted crash risk and the probability of that firm being a crash firm.

Concluding, it is important to mention that in finance, downside risk prediction is naturally a ranking problem. Investors care about knowing which companies are most exposed to losses, which ones should be monitored more closely and which ones are more likely to crash. Since exact numerical values are difficult to predict in financial markets, ranking can be a useful alternative if exact values are imperfect and difficult to obtain. RGA is a valuable metric that aligns with this decision issue, fitting even better than traditional metrics.

4.4 RGA and Subset Removal Analysis: Ranking Robustness

The baseline RGA provides one overall measure of ranking performance for this analysis. A shortcoming though is that the overall RGA does not reveal whether performance is evenly distributed across firms or not. Some companies could be easier to predict than others, since a high overall RGA could be driven by a subset of easy observations.

The purpose of the analysis in this section is to evaluate the model performance on progressively more difficult firms and assess the robustness of ranking quality after having removed the easiest cases.

The conceptual idea behind the analysis is that firms vary in prediction difficulty. Some companies can be forecasted very accurately, while others have a poor prediction. Through the removal of the easiest observations, it can be investigated how much of the ranking performance survives after the straightforward observations are removed.

This analysis is useful since the overall performance may overstate model quality. A stronger performance does not necessarily come from the whole data, it could originate from a small group of easy to rank firms. Through the evaluation, it is assessed whether the model remains effective even when these easy to predict firms are removed from the data set. This provides insights into the robustness of the ranking performance, since it's tests if the model captures the general structure or the data or only the obvious cases.

For the error calculation the first step is to obtain the model predictions $\hat{y}$ and compare them with the true outcomes y. The absolute prediction error is then computed for each observation

$$error_i = |y_i - \hat{y}_i|$$

The absolute error is used for the calculations since it measures the prediction quality and ignores the direction of error. This way the focus is on the magnitude of the deviation. The smaller the value of the error, the more accurate the prediction.

In the next step, firms are ranked by their prediction quality. Companies are sorted according to their absolute prediction errors, where the firms with the lowest error are first and the ones with the highest error last. The interpretation is that firms at the top of the ranking are the easiest once to predict and their prediction comes closest to their realised outcome. Contrary, firms at the bottom are the hardest to predict and their predictions are the furthest from the realised outcome.

The removal procedure is done in levels, going from 0% to 80% in 10% steps. For each removal level, firms are ranked by their prediction error. Firms with the lowest errors are removed and the remaining observations are retained. Afterwards, the RGA is recomputed on the remaining sample. It is important to mention that the removed firms are not random, they are the best predicted firms of the sample. The sample becomes increasingly difficult to predict, since the hardest firms are left.

After each removal step, a new data set is created and a new RGA is computed using only the remaining firms. The methodology to compute the new RGA is maintained, only the composition of the sample changes.

RGA is recomputed after each subset removal, since the objective is to measure a ranking quality under an increasing difficulty level. This allows direct comparison with the baseline RGA and it isolates the effect of removing easy observation from the sample.

The result of the gradual removal of companies is represented in the AURGA graph (area under RGA). The X axis represents the percentage of firms removed and the Y axis shows the RGA score computed on the remaining sample.

4.5 Noise Robustness Analysis: RGR

The purpose of this section is to test whether the model rankings are stable even when predictions are disturbed. The baseline RGA measures the ranking performance under clean predictions and

the noise robustness tests whether the ranking survives when uncertainty is introduced in the sample.

Financial predictions are naturally noisy and disturbed, so small perturbations should not completely change the ranking of a reliable model. If the performance of the ranking collapses after small noise, then this suggests that the model is fragile and that the predicted ordering is unstable. On the other hand, if the ranking performance remains similar even after having added noise this shows that the model is robust and that the ranking contains a strong structural signal.

This analysis matters especially in finance because financial data is noisy by nature. Model predictions are uncertain and investors are not able to observe perfect risk signals. This is why a useful model should keep its relative risk ordering even under small disturbances. The noise robustness analysis is especially relevant for downside risk since exact losses are difficult to predict and this table ranking is often more useful than singular predictions. Through the Robustness analysis it is checked whether the models ranking is reliable enough to be meaningful in predictions.

Giudici and Kolesnikov (2025) use an RGR module for robustness. Their SAFE AI compares the original predictions with the perturbed predictions using the following formula

$$RGR = 1 - \frac{CvM(pred, pred_{pert})}{G(pred)}$$

Through this analysis, Giudici and Kolesnikov (2025) evaluate how stable there are predictions are relative to their original distribution. Noise can be added through input features. This way, the original and perturbed predictions can be compared.

For this thesis the same robustness idea was maintained. Predictions are perturbed and then their stability is assessed, but contrary to Giudici and Kolesnikov (2025), for this analysis the SAFE RGR is not computed directly. Instead, the code recomputes the RGA after perturbing predictions. This leads to an RGR evaluation that is based on an RGA robustness analysis.

Giudici and Kolesnikov (2025) use separate metrics in their study. They implement RGA for ranking accuracy, RGR for robustness and RGE for explainability. Contrarily, in this thesis the RGA is used as the single common metric to compare the three trustworthy principles of machine learning.

In this section, the robustness of the model is measured by how much RGA changes after adding noise. This is the main methodological adaptation: robustness is expressed on the same scale as the baseline model performance. The advantage of this approach is that it allows a direct comparison with the baseline RGA, as well as a direct comparison across models and across outcomes. The name RGR is being kept since the analysis corresponds to the robustness analysis, only that its implementation is RGA based.

The function name is noise_robustness_rga(). The inputs for the function are the true outcomes values y, the out of fold predictions yhat, the noise level, the number of repetitions and the random seed. y and yhat are converted into numerical arrays and the standard deviation of predictions is calculated through

$$\sigma_{\hat{y}} = std(\hat{y})$$

For each noise level, a gaussian noise is generated and added to the predictions. Then the RGA is recomputed. This process is repeated 50 times, so that in the end 50 different RGA values are

obtained. From these 50 different RGA values, the mean RGA is calculated to get the average ranking performance after perturbation

$$\overline{RGA} = \frac{1}{N}\sum_{i=1}^{N} RGA_i$$

This means that each robustness estimate is based on 50 independent perturbation experiments which reduces the dependence on one lucky or unlucky noise draw. Through these calculations, the standard deviation measures how much these 50 RGA values vary around the average. A low standard deviation indicates that the model robustness is very stable an that different random noise draws give similar RGA values. On the contrary, a high standard deviation shows that the robustness result is more uncertain and that the different random noise draws affect the ranking performance substantially.

The code stores the mean RGA, the standard deviation of the RGA and the number of repetitions. With these indications, it computes the area under the robustness curve (AURGAR).

The noise is constructed in the following way. Noise is added to the predictions, not the input variables, and for each firm

$$\hat{y}_i^{noise} = \hat{y}_i + \epsilon_i$$

The random noise term is a gaussian noise

$$\epsilon_i \sim N(0, \sigma_{noise})$$

Gaussian noise is used since it's a standard noise in robustness testing. It is symmetric around 0 and does not systematically increase or decrease predictions. In addition, since its mean is equal to 0, it has no directional bias and achieves a controlled standard deviation. This leads to the gaussian noise being useful for testing.

The noise standard deviation is computed through

$$\sigma_{noise} = \lambda * std(\hat{y})$$

Lambda represents the noise level, which is scaled relative to the prediction variability. This makes the noise comparable across outcomes and models. If the predictions have a larger dispersion, that indicates that the absolute noise becomes larger. On the other hand, if predictions have smaller dispersions, the absolute noise becomes smaller.

For this analysis different noise levels have been used: 0%, 1%, 2%, 5%, 10%, 20%, 30% 40%, 50%. 0% represents the clean baseline prediction, a small noise indicates fine stability, while a medium noise shows moderate perturbation and a large noise analyses the stress test. The different noise levels are found on the X axis of the AURGA: they are the noise level in percent of the prediction standard deviation.

As priorly mentioned, the noise is random. This means that one noise drop could lead to an accidental increase or decrease. For this reason, the code repeats each noise level 50 times to avoid lucky or unlucky draws. For each repetition, a new noise vector is generated, noisy predictions are created and the RGA is calculated. After the repetitions, the average RGA is computed, as well as the standard deviation. The mean RGA is then the expected ranking performance under that noise level, while the standard deviation represents the uncertainty and variability of the robustness response. The lower the standard deviation is, the more stable the response. On the contrary, the higher the standard deviation, the more sensitive the response.

For the RGA recomputation, after each noisy prediction a vector is created

$$RGA(y, \hat{y}^{noise})$$

The true outcome stays unchanged and only the predictions are perturbed. Through this, it is directly tested whether the predicted ranking remains aligned with the realised downside risk.

The important difference from Giudici and Kolesnikov (2025) is that their RGR compares the original predictions to perturb predictions, while this thesis compares the perturbed predictions to the realised outcomes using RGA.

The graph shows the AURGAR (Are Under the RGA Robustness Curve). It is computed using on the X axis the normalized noise levels, while on the Y axis the mean RGA. A higher AURGAR means that the RGA remains high across the noise level, indicating that the model ranking is robust. On the other hand, a low AURGAR shows that the RGA drops quickly and that the model ranking is fragile. A flatter curve shows a more robust ranking, while a steeper downward curve indicates a less robust ranking. A shaded band is included in the graphs, which represents plus/ minus one standard deviation of RGA across the repetitions. A wider shaded band suggests greater instability across random noise draws, while a narrower shaded band represents a more consistent robustness behaviour. Through the AURGAR, the robustness curve can be summarised in one number.

This analysis is applied both to LASSO/ Logistic L1 with K=10 and Random Forest, where both models are tested for all three downside variables (maximum drawdown, downside volatility and crash indicator).

Out-of-fold predictions are used, to which the noise is added. The predictions are generated then on observations that are not used to train the model in that fold. This way the robustness evaluation is avoided on in sample fitted values. This step is important because noisy in sample predictions could overstate the stability of the model.

To make the comparison across models and outcomes easier, the same noise levels are used for all models, as well as the same number of repetitions and the same RGA based evaluations. This allows a fair comparison between LASSO K=10 and Random Forest, as well as continuous outcomes and crash outcome.

4.6 Mechanistic Interpretability: RGE

The baseline RGA evaluates the overall ranking performance, but it does not tell which variables generate the performance. The objective of this section is to understand the internal mechanism of the model and to identify the variables that are responsible for ranking the firms by downside risk. Through this part, the analysis moves from prediction toward explanations.

Giudici and Kolesnikov (2025) evaluate the sensitivity of the predictions when features are removed and they use a separate RGE metric. Instead, this thesis adopts their explainability philosophy, but instead of the SAFE RGE metric, this analysis recomputed the cross validated RGA after each feature removal. Here, the explainability is therefore measured through ranking degradation and this way all analysis remain on the same scale, namely RGA.

The core idea is to start from the complete model and progressively remove information. The model is retained, but every time information is removed, the cross validated RGA is recomputed.

Through this analysis it is observed how the ranking performance changes when variables are removed.

The interpretation logic behind this section is that if the model shows a large decrease in RGA while information is progressively removed, it signals that the variable discarded is important. The model relies heavily on that information and through the removal of that variable, the ranking performance deteriorates majorly. On the contrary, if there is a small RGA decrease after a variable is eliminated, it shows that it contributed little to the ranking. The information is likely redundant and other variables can compensate the missing data. In case of an RGA increase, it shows that they removed variable may introduce noise into the model. Keeping that variable in the model can reduce ranking performance and the model works better without it.

The retaining of the variables is important because coefficients are not simply deleted. The model is estimated again after each removal and the remaining variables are allowed to adjust. This way the model can reflect the true dependence of information.

This analysis is very important for the financial sector since it identifies the drivers of downside risk rankings. It can improve the transparency of variables and their contribution to the prediction. In addition, it facilitates the economic interpretation and it links machine learning outputs with the financial theory.

4.6.1 LASSO K = 10 Mechanistic Interpretability

LASSO is a suitable model for ablation, since it is a white-box model with explicit coefficient estimates. The variable selection is observable and it allows for a direct interpretation. LASSO K=10 is the main interpretation model for this thesis, even though LASSO K=5 is computed as well.

The variable ordering is based on three different points. First selection frequency, namely the number of cross validated folds in which the variable is selected. Second, the coefficient magnitude which implies that the larger the absolute coefficient is, the stronger its contribution is. Third, the dominant variable treatment, where the Volatility_2023 variable was placed at the beginning of the ablation sequence. Consequently, the code forces the volatility variable in first position and it is treated as a dominant predictor. This choice is it taken since volatility represents the primary financial risk proxy and is consistently identified among the most relevant predictors for model estimations in finance. This way before looking at other variables, the code first tests what happens when the most obvious risk variable is removed.

The model generates a different removal sequence for each of the three downside variables. Therefore, the model avoids assuming that the downside risks have the same variable importance structure.

As the first step, the whole model is computed. All variables are included and the baseline cross validated RGA is calculated. In the subsequent steps the highest ranked variable is removed and the feature matrix is rebuilt. The model is retained and the cross validated RGA is recomputed. In the next steps, the variables are removed cumulatively. This means that for example in step two the first two variables are removed and in step three the first three variables are removed. The cross validated RGA is recomputed after every variable elimination. Lastly, the worsening of the ranking can be controlled by monitoring the RGA drop

$$RGA_{baseline} - RGA_{current}$$

The same cross validation framework as the baseline model is used. For the regression outcomes the LASSO model is used, while for the crash outcomes the logistic regression with L1 penalty.

Through this procedure, the most important variables for the LASSO ranking performance can be identified. This step provides a transparent explanation for the model behaviour and allows to comprehend the outcome.

4.6.2 Random Forest Mechanistic Interpretability

Random Forest is a black-box model, so coefficient interpretation is not available. For this reason, an additional procedure is needed to analyse variable importance.

The variable ordering is based on the feature importance of the Random Forest, which is extracted from the trained data. In this section as well, the code uses a dominant variable treatment. Volatility_2023 is manually placed first, the same treatment as for LASSO.

In the first ablation step, the full Random Forest model is computed to obtain the baseline cross validated RGA. In the subsequent steps, the variable is removed, the predictor matrix is rebuilt, the model is retrained and the cross validated RGA is then recomputed.

Same as in LASSO, for Random Forest the variables are cumulatively removed through subsequent removal. Through this procedure, information is progressively reduced and the ranking degradation can be monitored.

As per the earlier model, the same cross validated framework is used for the Random Forest. After each variable is removed, the model is reoptimized instead of using the same parameters selected for the full model. This makes sure that any changes in the RGA is due to the loss of information from removing variables and not because of suboptimal model specification. Then, the out-of-sample predictions are generated and lastly the cross validated RGA is computed.

A large decline in RGA indicates that the Random Forest model strongly depends on the removed variables. On the other hand, a small decline in RGA shows that the Model is able to compensate through the remaining variables.

Contrary to LASSO, the importance in Random Forest is distributed across many variables. The result of this is that the effect that a variable has on a model is normally more diffused. As a consequence, the interpretation is less direct.

Through ablation, the black-box model is opened, allowing for the identification of the variables that cause the performance of the model. In addition, it enables the comparison of Random Forest with LASSO.

4.6.3 Comparison of LASSO and Random Forest Interpretability

Both models start from a full feature set, then remove variables sequentially and retain the model. Afterwards, they recompute the cross validated RGA and evaluate the ranking degradation.

Furthermore, they have use the same evaluation metric, namely RGA. This makes them directly comparable and formulates a consistent thesis methodology.

Since LASSO is a white-box model, its coefficients are observable and the variable selection is explicit. This makes the economic interpretation easier and more straightforward. Due to its sparse structure, fewer variables drive predictions and the dependence of the data is more concentrated.

On the other hand, Random Forest is a black-box model, meaning that its internal structure is not directly observable. The importance of variable is distributed meaning that the information is spread. This results in a more complex dependencies structure.

The consequences of these two different models is that there is an interpretability trade off. LASSO has a higher transparency but a lower flexibility. Random Forest has a higher flexibility but a lower transparency.

Analysing both model gives the analysis a robust benchmark and a comparison method. It verifies whether conclusions depend on the model type used in the evaluation.

5. Empirical Results

In this section, the empirical results will be analysed and evaluated. The comparison is done between LASSO K=10/ Logistic L1 K=10 and Random Forest for all three downside outcomes (Maximum Drawdown, Downside Volatility and the Crash Indicator).

As mentioned priorly in the methodology section, preliminary analyses were computed for LASSO, where the model was evaluated for two different specifications: K=5 and K=10. In the first specification, the model is constrained to retain approximately five predictors, while in the second one the model is constrained to retain approximately ten. The predictive information was significantly higher for the LASSO model that was retaining more predictors, since the model has More variables to base its forecasting on. The same procedure was applied for Logistic L1, achieving the same results as for LASSO. Thus, for both LASSO and Logistic L1, K=10 was chosen as the underlining algorithm for linear models for the whole thesis.

Since maximum drawdown and downside relativity are continuous outcomes, the LASSO model is used. On the contrary, since the crash indicator is a binary variable, LASSO is substituted by the Logistic L1 model.

5.1 Baseline Ranking Performance

The first empirical evaluation of model performance of this thesis is about the baseline RGA. It analyses how well models rank firms based on the future downside risk. The performance is assessed with a cross validated RGA and the crash indicator is evaluated through an F1 score.

Outcome	LASSO / Logistic L1 K=10	Random Forest
MaxDrawdown	0.8006	0.8090
DownsideVol	0.8127	0.7964
Crash	0.6631	0.6852

5.1.1 Maximum Drawdown

Analysing the results for maximum drawdown, it is observable that the LASSO RGA is equal to 0.8006, while the Random Forest RGA corresponds to 0.8090. Both exceed 0.8, which is overall a very high-ranking performance. The difference between the two models is extremely small, being around 0.0084, where Random Forest has the marginally higher result.

Both models are able to successfully rank firms based on the future maximum drawdown. Their high scores show that the ordering of the firms is captured well and that the predictive signal is strong. It is worth to mention, that a little more benefit is obtained by moving from a linear model (LASSO) to a nonlinear model (Random Forest). The maximum drawdown seems to be predictable using the available variables. Both LASSO and Random Forest obtain almost the identical ranking quality.

Concluding, it can be said that the different performance gap between the two models is negligible, since the advantage that Random Forest gives is very small. This result suggests that

both models are able to extract similar information from the available data. The complexity of using a nonlinear model does not generate a substantial improvement compared to the linear model.

5.1.2 Downside Volatility

Evaluating the downside volatility result, the LASSO RGA corresponds to 0.8127, while the Random Forest RGA is equal to 0.7964. Here the LASSO RGA is the highest baseline RGA observed among all outcomes. Similarly, as for maximum drawdown, both models for downside volatility have an RGA around 0.8.

This outcome tells that the downside volatility is the easiest outcome for the model to rank. It indicates a strong relationship between the variables and the future downside volatility. The ranking quality is high end contrarily to maximum drawdown, LASSO outperforms Random Forest. This suggests that the linear structure of LASSO is able to capture most of the predictive information.

Comparing the two model outcomes, there is a very small difference between LASSO and Random Forest. LASSO has an advantage of 0.0163. It is the opposite of maximum drawdown, since for the downside volatility, LASSO is performing slightly better. Overall, the difference is minimal and the performance is essentially equal. This result supports the robustness of the findings across different models.

5.1.3 Crash Indicator

Since the crash indicator is a binary variable, the Logistic L1 model is used. Its RGA equals 0.6631, while the Random Forest RGA is 0.6852. This result is lower than the previous continuous outcomes, which indicates that the ranking task here is substantially more difficult.

The interpretation behind the numbers is that the crash indicator is harder to predict than the maximum drawdown and downside volatility. A reason for this is that binary outcomes contain less information and this results in an event prediction which is more challenging. As a result, a lower ranking accuracy can be expected. Additionally, the Random Forest model shows a noticeable improvement over the Logistic L1 model.

The reason for why the Crash Indicator results in a lower RGA than the previous continuous outcomes is, that the sample contains only 42 crash firms, since it's constructed with 10% of the companies. This result in an imbalanced outcome. Furthermore, classification problems are inherently harder since less information is available compared to continuous outcomes. As a result, extreme events are more difficult for the model to identify.

As it can be seen confronting the RGAs, Random Forest outperforms Logistic L1. This indicates that the crash indicator benefits from non-linear models and more model complexity. This result difference is the largest one observed among the three outcomes, being of 0.0221.

5.1.4 Overall Comparison Across Outcomes

Comparing the overall ranking of outcomes, it can be stated that the downside volatility has the highest RGA. Maximum drawdown is a close second while the crash indicator is substantially

lower. This indicates that continuous outcomes are easier for the model to rank since they contain richer information. Contrarily, binary outcomes are more difficult to predict and the ranking performance decreases.

It is important to note that for the continuous outcomes both LASSO and Random Forest are similar. The advantage of a nonlinear model, namely Random Forest, is noticeable only for the crash outcome.

It can be concluded that additional model complexity provides only limited overall gains.

5.2 Subset Removal Analysis (AURGA)

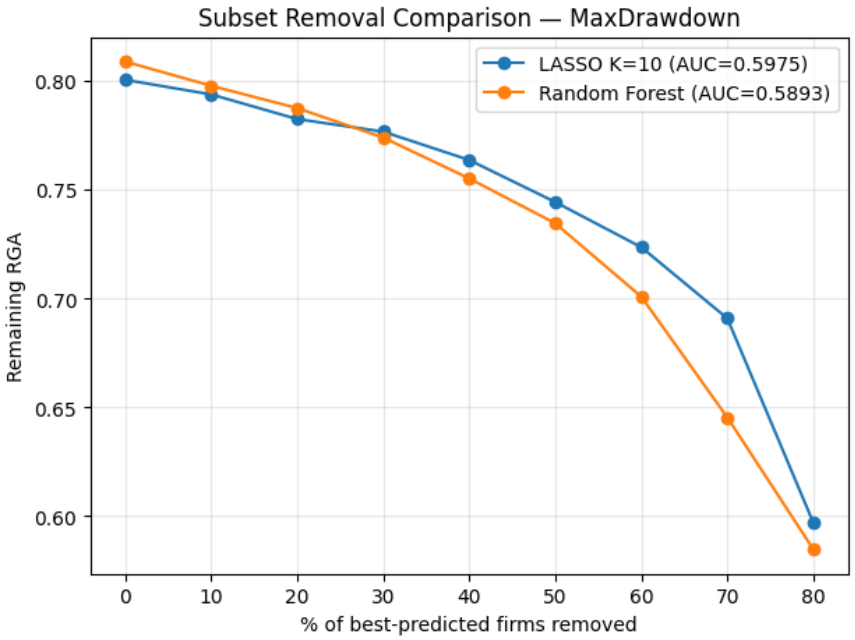
The purpose of this section is to analyse if the baseline ranking performance maintains its stability, even when the easiest observations are removed from the sample. While the baseline RGA shows the total ranking performance of the full sample, the subset removal test if the performance mainly depends on firms that are easy to predict. To evaluate this, the companies that have the lowest absolute prediction error are removed first from the sample. This way, the remaining firms get progressively harder to rank. Furthermore, RGA is recomputed after each removal level.

This section focuses on the empirical comparison between the LASSO K=10/ Logistic L1 K=10 and the Random Forest model.

Outcome	Model	AURGA	RGA 10%	RGA 40%	RGA 60%	RGA 80%
MaxDrawdown	LASSO K=10	0.5975	0.7939	0.7638	0.7237	0.5970
MaxDrawdown	RF	0.5893	0.7979	0.7553	0.7009	0.5848
DownsideVol	LASSO K=10	0.6064	0.8061	0.7708	0.7315	0.6320
DownsideVol	RF	0.5735	0.7841	0.7360	0.6845	0.5252
Crash	Logistic L1 K=10	0.4878	0.6589	0.6359	0.5947	0.3872
Crash	RF	0.4899	0.6812	0.6597	0.5891	0.2897

On the X axis the percentage of the best predicted firms that are removed is depicted while on the Y axis the remaining RGA is shown.

5.2.1 Maximum Drawdown Subset Removal



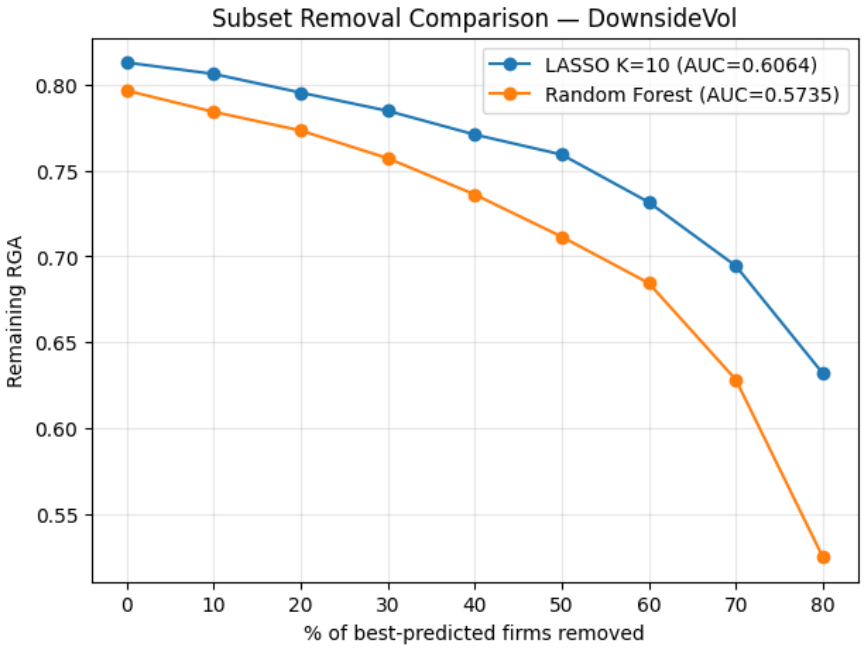
LASSO K=10 tarts from the full sample RGA which corresponds to 0.8006. As seen in the graph, the RGA decreases gradually as the best predicted firms are removed. After the 10% removal, the RGA is at 0.7939. After 40% removal, it is at 0.7638. After 60% removal, the score is at 0.7237 and after 80% it is at 0.5970. Additionally, the area under the RGA curve (AURGA) equals 0.5975. From these numbers it can be concluded that the decline is gradual up until the 60% mark. Afterwards, a stronger drop appears due to high removal levels. Overall, the model still gives a meaningful performance ranking even after almost all the easy cases are removed.

For the Random Forest, the initial performance off the full sample RGA is slightly higher than LASSO, being at 0.8090. The RGA declines also progressively. At 10% removal it corresponds to 0.7979, after 40% it equates 0.7553, after 60% it is 0.7009 and after 80% it reads 0.5848. These numbers indicate that the decline becomes stronger after 50 to 60%.

Comparing both models, it can be said that their curves are shaped in a similar way. While Random Forest starts slightly above LASSO, after the 20% removal mark the positions become inverted and LASSO becomes more robust at higher removal levels. At 80% removal, LASSO is equal 0.5970 and Random Forest corresponds to 0.5848. Random Forest's final RGA at 80% is 0.0122 lower than LASSO's.

It can be concluded that Random Forest is marginally better on the full sample, but that LASSO is more stable in ranking the remaining observations that are more difficult. So LASSO retains more RGA after removing the easiest to predict firms.

5.2.2 Downside Volatility Subset Removal



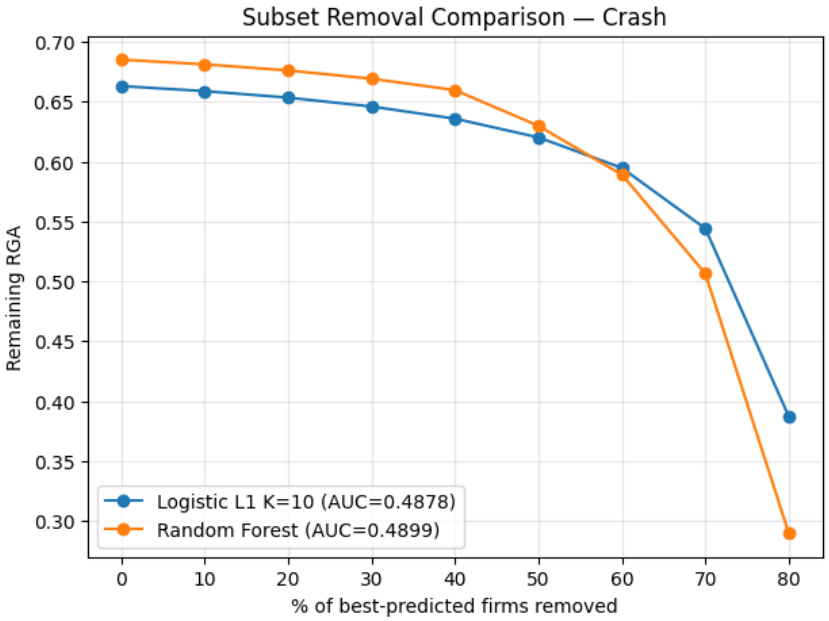
The full sample RGA for LASSO K=10 is equal to 0.8127, which is the highest starting RGA among all three outcomes. As can be seen in the graph, the decline is smooth and gradual. After 10% removal the RGA is 0.8061, after 40% it is equal to 0.7708, after 60% it corresponds to 0.7315 and after 80% removal it is at 0.6320. The graph shows that the curve remains strong through most levels an even after the 80% removal mark, the RGA is above 0.63. Furthermore, the AURGA is equal to 0.6064. These results indicate the strongest persistence in analysing the harder cases between all three downside measures.

The full sample RGA for Random Forest starts on the other hand at 0.7964 and its RGA declines faster than the one of LASSO. After 10% removal it is at 0.7841, after 40% it equates 0.7360, after 60% it corresponds to 0.6845 and after 80% it is at 0.5252. From the graph it can be seen that the Random Forest curve remains close to the LASSO curve early on and that the gap becomes bigger the higher that removal level becomes. The final remaining Random Forest RGA is below LASSO's by over 0.1.

Comparing the two models, LASSO outperforms Random Forest throughout all of the removal path. The difference becomes more highlighted as more easy firms are eliminated from the sample. At 80% removal, LASSO corresponds to 0.6320 while Random Forest is 0.5252. The difference is of 0.1068 in favour of LASSO, which is the largest advantage between subset removal continuous outcomes.

Concluding, it can be said that LASSO is more persistent at ranking the hard cases of downside volatility, while Random Forest Loses more performance when ranking difficult cases.

5.2.3 Crash Indicator Subset Removal



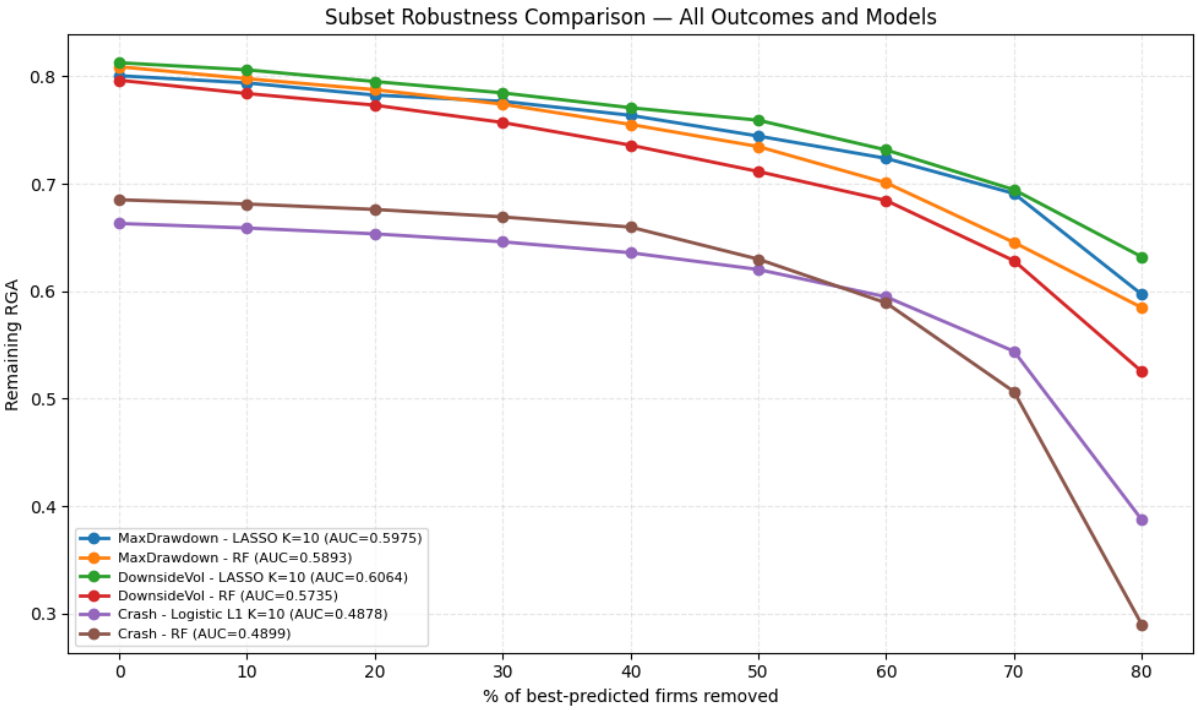
For Logistic L1 K=10, the full sample RGA corresponds to 0.6631. This is a lower starting RGA then for the continuous outcomes. The RGA then declines slow at first. After 10% removal the RGA is at 0.6589, after 40% it corresponds to 0.6359 and at 60% it is at 0.5947. There is a strong decline in the RGA after the 70 and 80% removal mark, as can be seen since the RGA at 80% is 0.3872. Here, that AURGA corresponds to 0.4878. The graph shows that the curve deteriorates sharply at the late stage which shows that the crash ranking becomes very weak once the hardest observations are the only ones remaining.

The Random Forest start from the full sample RGA at 0.6852. This is a higher starting point than for the Logistic L1 model. The RGA remains relatively high at the early removal stage: after 10% removal it is at 0.6812, after 40% it shows 0.6597 and after 60% it corresponds to 0.5891. Like for Logistic L1, Random Forest presents a sharp drop after the 70 and 80% removal mark, since the 80% RGA equals 0.2897. The AURGA is 0.4899. The Random Forest model for the crash indicator presents the strongest late stage collapse among all the curves. It performs best initially but loses more performance at extreme removal levels.

At the beginning of the curve, Random Forest start clearly above Logistic L1 and it maintains the advantage through all the earlier removal levels. After the 50% removal mark is hit, Random Forest goes below Logistic L1 and deteriorates strongly. At 80% removal, Logistic L1 is at 0.3872 compared to Random Forest's 0.2897. While Random Forest starts above Logistic L1, they invert places and Logistic L1 ends above Random Forest after.

Concluding it can be said that Random Forest is able to rank the crash risk better on the full sample, while Logistic L1 presents more stability for ranking the most difficult subsets. The outcome of the crash indicator shows the largest degradation of subset removal between the three downside measures. For this reason, it can be described as less stable than the continuous downside outcomes.

5.2.4 Overall Comparison Across Outcomes



All curves on the three models decline when the easy to predict firms are removed from the sample. This is an expected result since the remaining firms are harder to rank. The decline of the score when removing easy confirms, that is contribute to the higher baseline RGA values. Overall, no model is unaffected by the removal.

The continuous outcomes maximum drawdown and downside volatility show a much smoother decline. They are both able to retain a stronger RGA even after large removals: at 80% removal, LASSO is above 0.59 and Random Forest remains above 0.53. Continuous downside risk measures have a higher robustness to the removal of easy to predict firms than the crash indicator.

On the opposite, the crash already starts at a lower baseline RGA and the decline in the late stages is much more sharply. The final RGA is lower as well, corresponding to 0.3872 for Logistic L1 and to 0.2897 for Random Forest. When removing the easy observations from the sample and leaving only the hard is to predict once, the crash ranking becomes more fragile. Overall, the crash indicator is the hardest outcome to predict under subset removal.

From both the maximum drawdown and the crash indicator it can be seen that Random Forest often starts from a higher baseline RGA then LASSO/ Logistic L1. The latter though maintains more performance at the 80% removal mark for all outcomes. From this it can be concluded that LASSO is favoured in hard case robustness and that Random Forest is more dependent on easy observations.

The baseline performance suggests that both Random Forest and LASSO obtain similar results when applying the model to the full sample. What makes a difference is the subset removal. Here, LASSO is more robust when handling hard cases. This is especially applicable for the downside volatility and the crash indicator when the removal levels rage high percentages.

5.3 Noise Robustness Analysis (AURGAR)

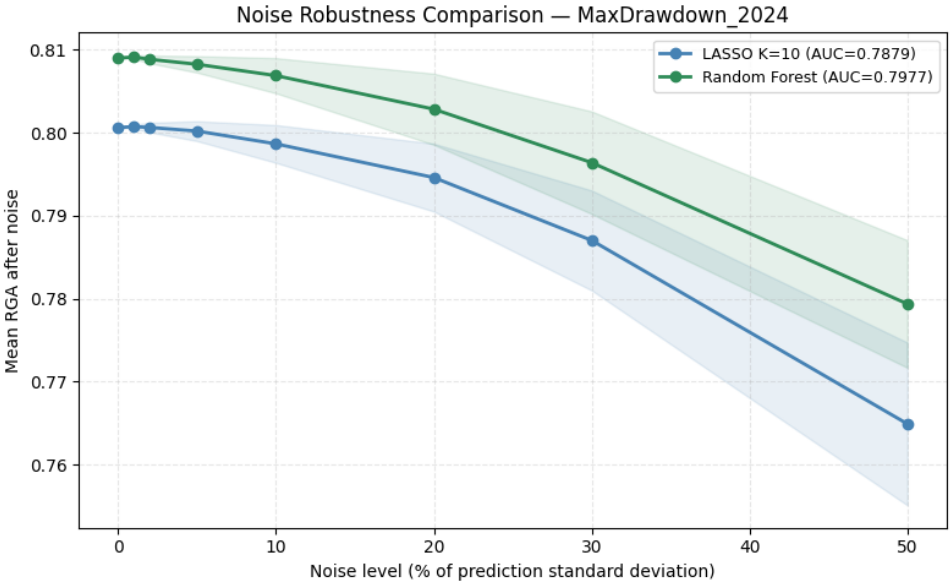
The purpose of this section is to evaluate whether the model rankings remain stable even after predictions are perturbed. Noise is added to the out of fold predictions and then the RGA is recomputed after every perturbation. Each noise level is repeated 50 times and then the average RGA is computed. The results are evaluated and summarised using the mean RGA, the standard deviation of RGA and the area under the RGA curve (AURGAR). The shaded region in the following graphs is the standard deviation, that helps to evaluate how sensitive the result is to the random noise draw, since it shows how much the 50 RGA values differ from each other. A small standard deviation means that every simulation gives almost the same result and that the robustness estimate is reliable. Contrarily, a large standard deviation shows that some simulations produce a much lower or higher RGA, so the robustness is less certain and the variability is higher. It mainly shows how much the observations diverge from the average.

Like in every section, LASSO K=10/ Logistic L1 K=10 is compared to Random Forest. The analysis is done for all three downside outcomes.

The following table represents the AURGAR values.

Outcome	LASSO / Logistic L1 K=10	Random Forest
MaxDrawdown	0.7879	0.7977
DownsideVol	0.8003	0.7845
Crash	0.6594	0.6812

5.3.1 Maximum Drawdown Noise Robustness



LASSO K=10 starts with the baseline RGA at 0.8006, which contains 0% noise. It is observable that under a very small noise percentage, the RGA remains almost unchanged, just a very small decline occurs. At 1% noise the RGA is namely 0.8007, at 5% a corresponds to 0.8002 and a 10% it is 0.7989. Increasing the noise percentage the RGA starts decline more. It is 0.7942 a 20% noise, 0.7870 at 30% noise and 0.7649 at 50% noise. This shows that the decline becomes more noticeable at high noise levels. In addition, the AURGAR corresponds to 0.7879. Furthermore, it

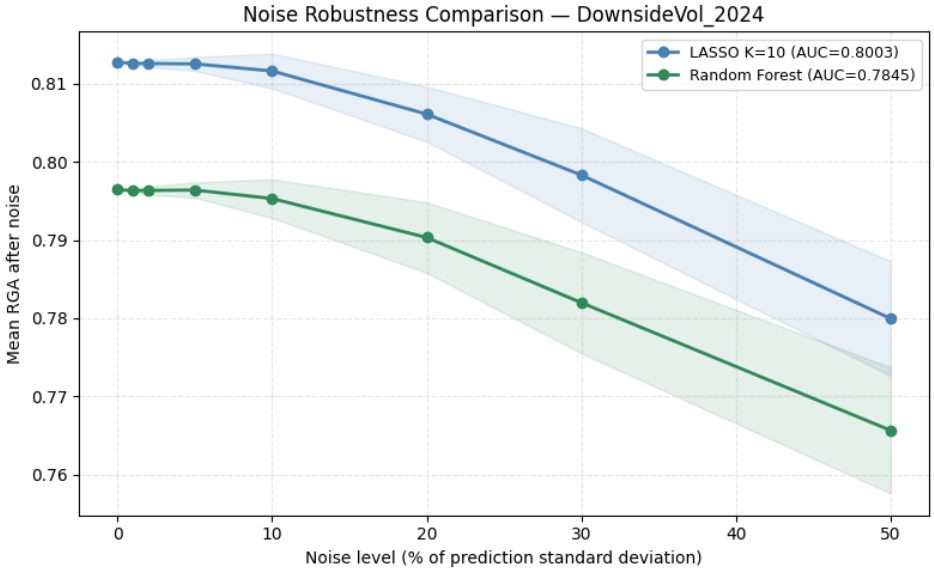
is observable that the standard deviation increases with noise: the standard deviation is 0 at the baseline RGA and 0.0091 at 50% noise. From this it can be taken that for the maximum drawdown, LASSO is stable if the perturbations are low and moderate, but that stronger noises reduce the accuracy and the ranking performance.

Conversely, Random Forest starts with a baseline RGA of 0.8090, which includes 0% of noise. From this we can affirm that Random Forest starts at the slightly higher RGA level then LASSO. Adding noise to the model the RGA starts to decline. At 1% noise, the RGA corresponds to 0.8091, at 5% noise it is at 0.8082 and at 10% noise it reaches 0.8069. Increasing the perturbation, an RGA of 0.8028 is reach at 20% noise, 0.7964 at 30% noise and 0.7794 at 50% noise. Moreover, the AURGAR is equal to 0.7977 and the standard deviation at 50% noise is 0.0080. From this evaluation it can be concluded that the Random Forest model ranking remains highly stable across noise levels, even with high perturbation percentages. The model is able to retain a higher RGA than LASSO when reaching high levels of disturbances.

Comparing the two models, they both show a very similar robustness curve. It has to be noted though, that Random Forest has a slightly higher AURGA, since it corresponds to 0.7977 compared to a 0.7879 of LASSO. The difference is about 0.0098. Across the whole curve and so for all noise levels, Random Forest remains slightly higher than LASSO. Both models though remain above 0.76 even at a 50% noise level.

Empirically, it can be concluded that maximum drawdown rankings are robust to prediction noise for both LASSO and Random Forest, even though the latter has a slightly higher robust outcome.

5.3.2 Downside Volatility Noise Robustness



For LASSO K=10, the baseline RGA at 0% noise is 0.8127. As mentioned in chapter 5.1, this is the highest starting RGA among all outcomes. Like in the prior section, the ranking remains very stable when the noise level is low. At 1% noise the RGA corresponds to 0.8126, at 5% it is at 0.8125 and a 10% it reads 0.8116. Increasing the noise percentage, at 20% the RGA is 0.8061, at 30% it shows 0.7983 and at 50% it equals to 0.7800. For LASSO K=10, the AURGA is 0.8003 and the standard deviation at 50% perturbation is 0.0099. From this analysis it can be interpreted that

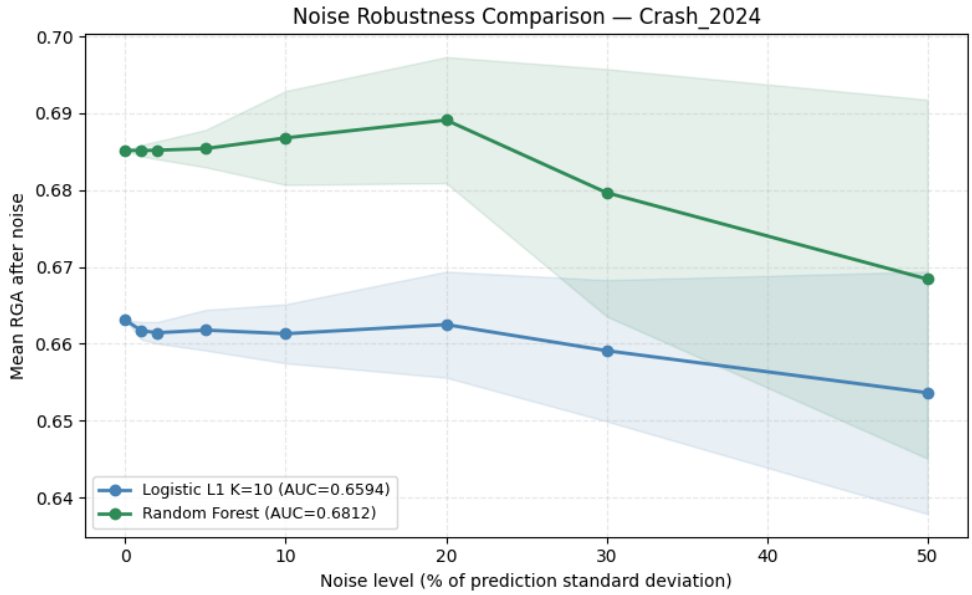
the downside volatility rankings of LASSO are very stable, especially at low noise levels, but even with larger perturbations the performance remains strong.

The second model, Random Forest, starts with a baseline RGA of 0.7964 with 0% noise. This is a starting point that is slightly below the one of LASSO. At 1% noise, the Random Forest RGA corresponds to 0.7963, at 5% noise it is at 0.7964 and at 10% noise it equals 0.7953. Continuing increasing the perturbation, at 20% the RGA is 0.7903, at 30% it is at 0.7820 and at 50% it corresponds to 0.7657. Additionally, the AURGAR is 0.7845 and the standard deviation reads 0.0102 at 50% noise. This evaluation shows that the Random Forest model remains robust as well, even if it's decline is slightly more accentuated than the one of LASSO.

Comparing the two graphs, LASSO outperforms Random Forest throughout the whole noise curve. Additionally, LASSO's AURGAR is at 0.8003 compared to Random Forests 0.7845. This shows a difference of 0.0158 in favour of LASSO. Both models are able to remain relatively stable under low and moderate noise levels. At 50% noise, LASSO corresponds to 0.7800 and Random Forest to 0.7657.

The empirical conclusion is that for the downside volatility, LASSO is the most stable model based on outcome, since it shows a slightly higher robustness then Random Forest, as well as a higher AURGAR.

5.3.3 Crash Indicator Noise Robustness



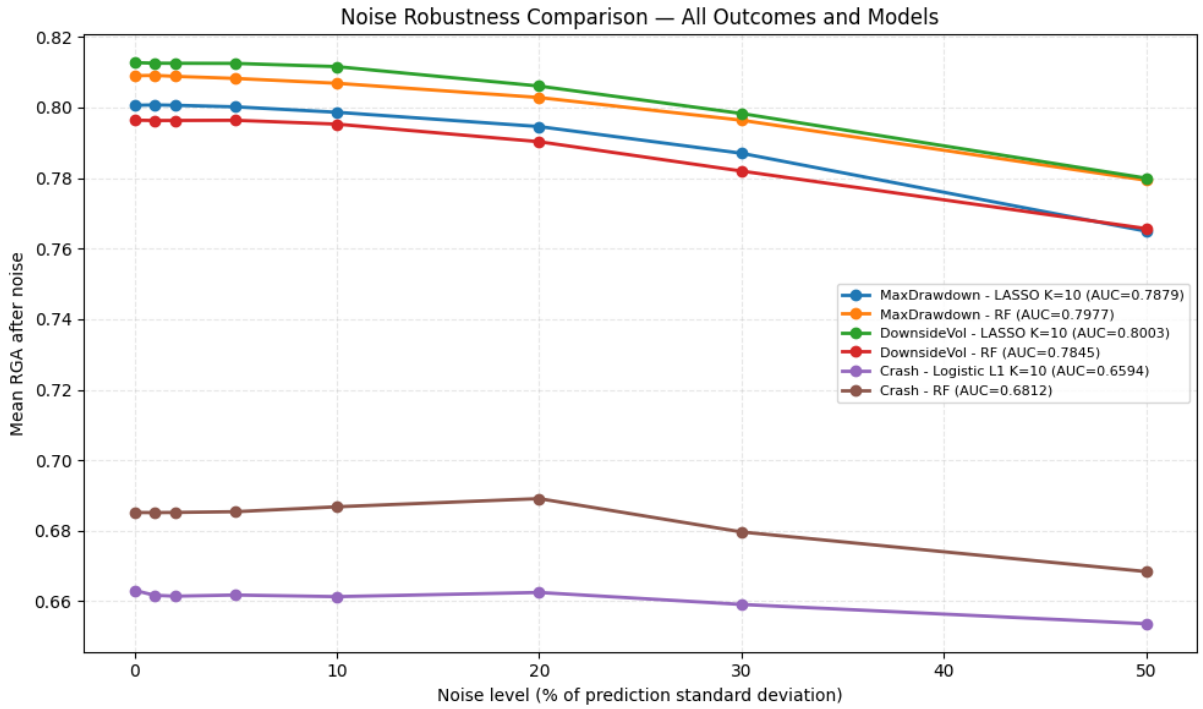
For the Logistic L1 K=10 the baseline RGA with 0% noise starts at 0.6631. This is noticeably a lower starting RGA then for the continuous outcomes. The mean RGA then remains around 0.66 even when reaching higher noise levels. At 1% and 5% noise, the RGA is at 0.6617 and at 10% it corresponds to 0.6613. Increasing the perturbation level, at 20% it is at 0.6625, at 30% it reads 0.6591 and at 50% it shows 0.6536. Furthermore, the AURGAR is equal to 0.6594 and the standard deviation increases strongly with the noise. At 10% it is namely 0.0043, while at 50% it is 0.0142. Evaluating this, it can be said that the average crash RGA is almost flat and around the same mean, although the uncertainty band widens when increasing the noise level. These results suggest that the average ranking is stable even when increasing perturbations, even though the model presents a higher variability when confronted with more noise.

Continuing with the Random Forest model, the baseline RGA at 0% noise is 0.6852. This start is higher than the one of Logistic L1. At 1% noise level the RGA shows 0.6851, at 5% it is 0.6854 and a 10% it reads 0.6868. Increasing then even more the perturbations, the RGA is 0.6891 at 20%, 0.6797 at 30% and 0.6684 at 50%. Moreover, the AURGAR is 0.6812 and the standard deviation increases like for Logistic L1 with the noise. It is at 0.0060 for 10% and 0.0158 for 50%. This analysis shows that the Random Forest crash ranking remains more robust and that the mean RGA stays higher than the Logistic L1 throughout all of the graph. Similar as for Logistic L1, the variability of Random Forest increases as well when the noise levels get higher.

Comparing the two models it is clear that Random Forest dominates the Logistic L1 AURGAR, since Random Forest has an area under the curve of 0.6812 while Logistic L1 has it of 0.6594. The difference is of 0.0218. Random Forest is able to keep a higher crash ranking performance throughout all the noise levels. Even though the Logistic L1 curve is flatter than the Random Forest one, it shows a lower performance level.

The empirical conclusion is that Random Forest is a more robust model for crash rankings in terms of the area under the graph. Overall, the crash ranking predictability remains lower for both models then it was priorly observed for the continuous outcomes.

5.3.4 Overall Comparison Across Outcomes



The general pattern that can be observed across all three downside risk measures is that all models are stable under small noise levels. The RGA changes are very small between 0% and 10% noise. The stronger declines are mostly noticeable at 30% and 50% noise. Moreover, the standard deviation increases with the augmentation in the noise levels. These wider bands at high noise levels are expected though, since perturbations become stronger and for this reason variability increases.

Comparing the continues outcomes to the crash outcome, maximum drawdown and downside volatility remain highly robust. Their AURGAR values are around 0.79. They both maintain strong

RGA's even at very high noise levels and this shows that they are more robust than the crash indicator. The latter on the other hand has a lower AURGAR for both models, ranging from 0.66 to 0.68. This shows that the crash remains more difficult than the continuous outcomes to estimate.

Comparing the three downside risk measures, for maximum drawdown Random Forest is the slightly more robust model, for downside volatility LASSO is the slightly better model and for the crash indicator the Random Forest is the more robust one. From this it can be observed that no single model dominates all three outcomes, the robustness ranking depends on the downside measure.

The main takeaways are that these results confirm that the model rankings are generally lower when perturbations are added to the model. Overall, the noise robustness does not overrule the baseline findings, it simply reinforces that the continuous downside risk is easier and more stable to rank than the crash risk.

5.4 Mechanistic Interpretability (AURGAE)

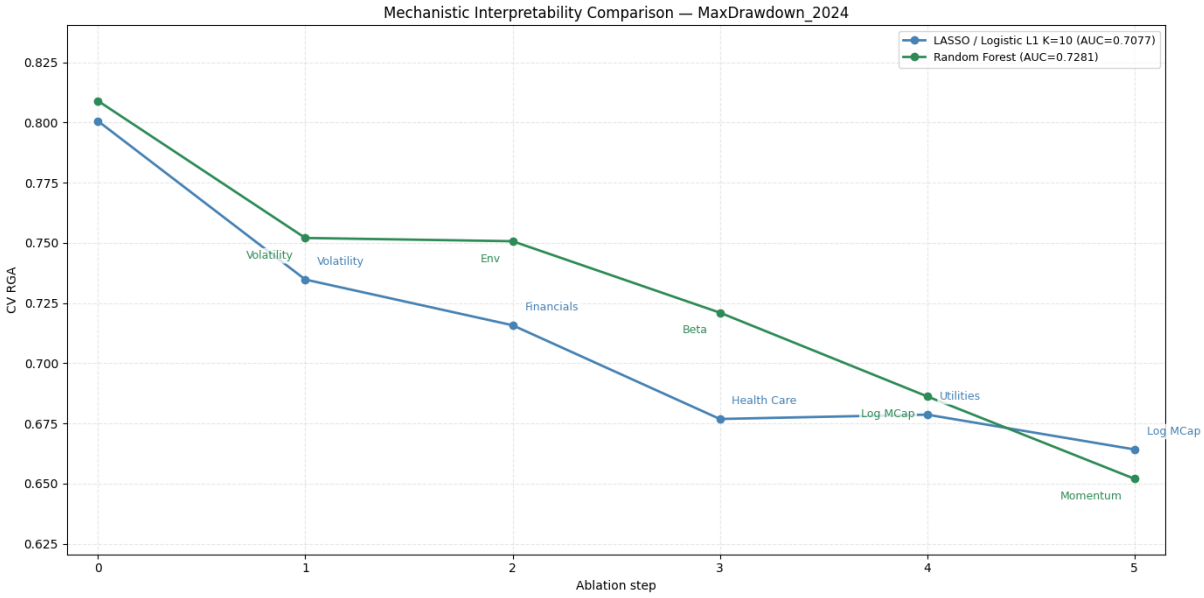
The purpose of this section is to report the empirical results that were obtained in the feature ablation analysis. It is evaluated how the ranking performance changes, when variables are sequentially removed from the sample. The cross validated RGA is used as the measure for the performance after each ablation step. As always, LASSO K=10/ Logistic L1 K=10 is compared to Random Forest for all three downside risk measures. The model is analysed by eliminating the five most useful variables for prediction. The area under the RGA curve (AURGAE) summarises this five-step ablation curve. The higher the AURGAE is, the more the model is able to keep the ranking performance even after variable removal. On the other hand, if the AURGAE is low, the model ranking performance deteriorates more quickly.

It is important to mention that in the ablation analysis volatility 2023 was placed first in the sequence to be removed. This choice is based on the theoretical relevance of volatility and it's strong significance in a preliminary feature importance analysis. All the remaining variables in the ablation sequence are chosen by the model.

Outcome	Model	AURGE	Baseline RGA	RGA after 5 removals	Total Drop
MaxDrawdown	LASSO K=10	0.7077	0.8006	0.6642	0.1364
MaxDrawdown	RF	0.7281	0.8090	0.6520	0.1570
DownsideVol	LASSO K=10	0.7380	0.8127	0.7168	0.0959
DownsideVol	RF	0.7188	0.7964	0.6516	0.1448
Crash	Logistic L1 K=10	0.6025	0.6631	0.6063	0.0568
Crash	RF	0.6019	0.6852	0.5237	0.1615

Each of the following graph will show on the X axis the ablation step and on the Y axis the cross validated RGA.

5.4.1 Maximum Drawdown Mechanistic Interpretability



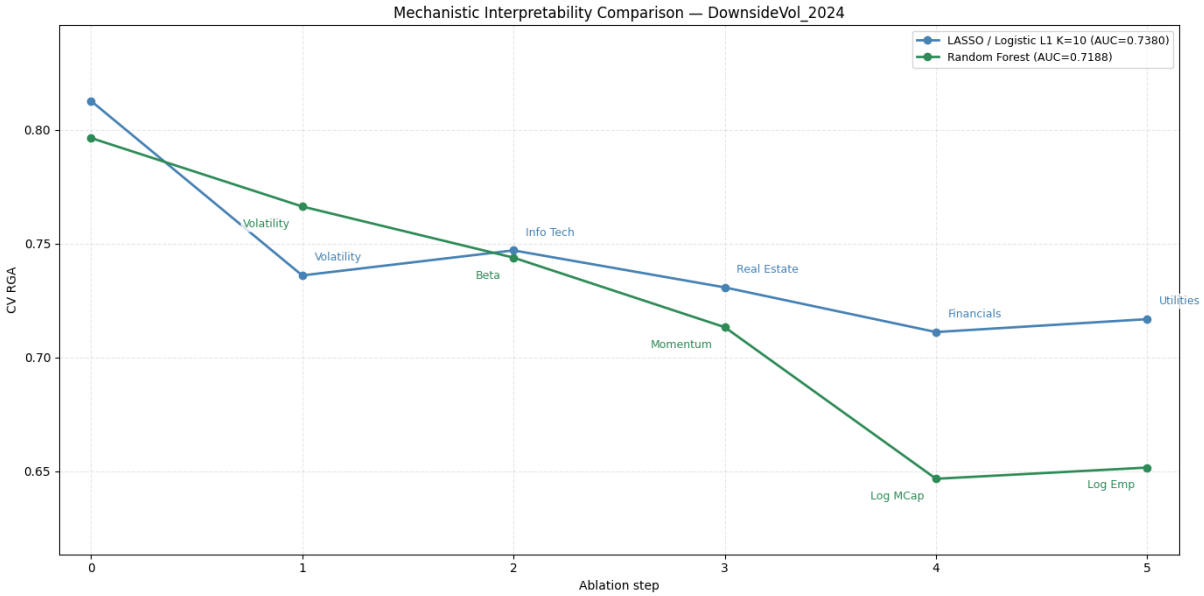
For LASSO K=10, the baseline cross validated RGA is 0.8006. After removing the first variable, namely the volatility in 2023, the RGA is at 0.7348. The next variable removed by the model is the Financials sector, which brings the cross validated RGA 0.7158. Next, the Health Care sector is removed, which brings the model to an RGA of 0.6768. Subsequently, the Utilities sector is deleted, which slightly increases the RGA to 0.6787. After removing the fifth variable, namely the logarithm of the market capitalization, the cross validated RGA drops to 0.6642. The total drop from the baseline RGA to the removal of the fifth variable is a decrease in 0.1364. This result shows that the curve falls sharply at the beginning and then stabilises after the removal of the third variable, becoming flatter. Additionally, the AURGAE is of 0.7077. The empirical message that comes through is that the LASSO drawdown ranking depends a lot on the variables that are removed in the first stages. After these initial variables are removed, the further variable eliminations have smaller effects on the model.

Looking at the Random Forest model, the baseline cross validated RGA is 0.8090. After removing the first variable, volatility 2023, the cross validated RGA drops to 0.7521. Then, the model removes the environmental score and the RGA falls to 0.7507. After the elimination of the third variable, namely the market beta, the RGA drops to 0.7210. Next, the logarithm of the market cap is removed, which brings the RGA to 0.6862. The final and 5th variable removed is the momentum of 2023 which brings the RGA to a final number of 0.6520. The total drop from the baseline RGA to the RGA after the fifth variable is removed, is a decrease of 0.1570. The AURGAE of Random Forest for downside volatility corresponds to 0.7281. From the graph we can observe that the Random Forest curve declines more continuously and that it ends at a lower RGA compared to LASSO. The main empirical message is that even though the Random Forest starts at a slightly higher baseline RGA, it loses more total RGA after variable removal.

Comparing the two models, it is observable, that they both start at a baseline RGA of around 0.8, even though Random Forest is slightly higher. After the fifth and final ablation, LASSO's RGA is at 0.6642 versus Random Forests 0.6520. This comparison shows that LASSO is able to retain more RGA after ablation than Random Forest, since the latter appears to be more sensitive to the removal of higher ranked variables.

The empirical result is that both models depend on specific key variables, but that Random Forest loses more ranking performance when features are removed.

5.4.2 Downside Volatility Mechanistic Interpretability



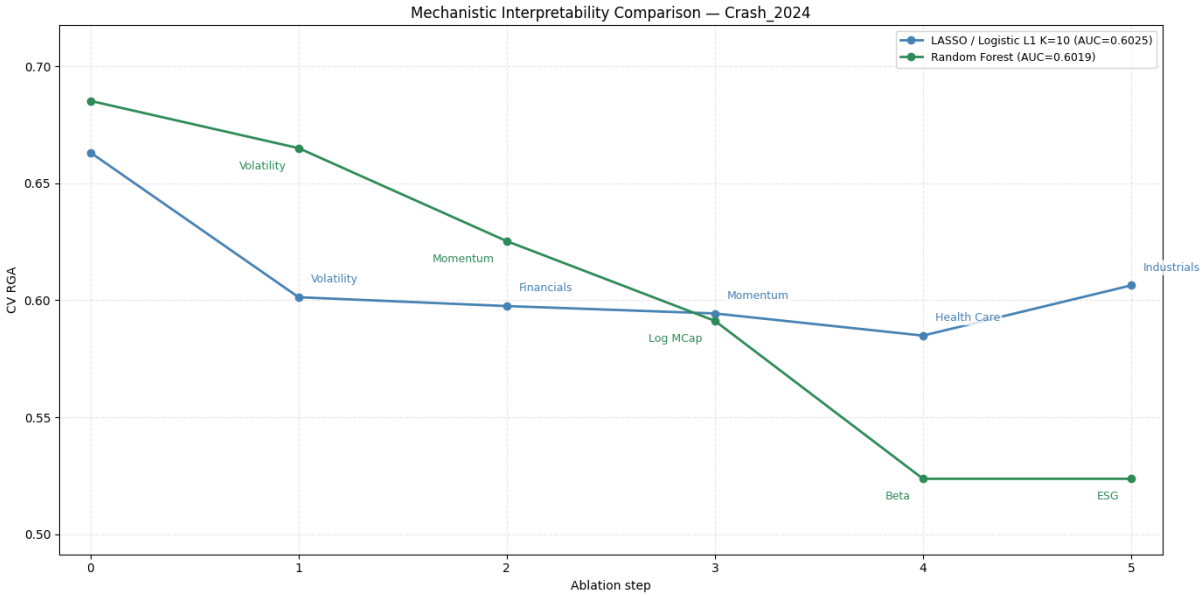
The downside volatility ablation measurement for LASSO starts with a baseline RGA of 0.8127. As always, the first variable removed is volatility, which leads the model to an RGA of 0.7360. The next variable removed is the information technology sector. Here the RGA partially recovers reaching 0.7470. The third variable removed is a Real Estate sector which leads to an RGA of 0.7308. Then the Financials sector is deleted, bringing the model to 0.7111. The fifth and last variable removed is the Utilities sector, leading the analysis to a slightly higher RGA of 0.7168. The overall drop from the baseline RGA to the RGA after the removal of the fifth variable corresponds to 0.0959. LASSO has an AURGAE of 0.7380. The curve shows an initial strong drop after the removal of volatility and then there is a more gradual decline. The final RGA remains relatively high though. Empirically, the downside volatility with LASSO is affected by the variable removal, but it overall remains relatively stable.

Random Forest on the other hand starts with a baseline cross validated RGA of 0.7964. After removing volatility, the RGA drops to 0.7663. The next variable removed is the market beta with which the RGA drops to 0.7438. Then the momentum 2023 gets removed which leads to an RGA of 0.7133. The fourth variable removed is the logarithm of the market capitalization, which brings the RGA to 0.6467 and lastly, the logarithm of the number of employees is removed, which brings the model to RGA of 0.6516. From the baseline RGA to the RGA after removing the fifth variable, the total drop corresponds to 0.1448. Here the AURGAE corresponds to 0.7188. Overall, the decline of Random Forest is much more accentuated than LASSO, since its final RGA is considerably lower. This shows that the ranking of the downside volatility for Random Forest is much more sensitive to cumulative feature removal than LASSO.

Comparing the two models, LASSO starts with a higher baseline RGA then Random Forest. The total drop in RGA of LASSO corresponds to 0.0959, while for Random Forest it is equal to 0.1448. This shows that LASSO retains more performance even when removing variables from the model. For downside volatility, LASSO has a clear advantage in ablation.

Empirically it can be concluded that both models Rd main useful after variable removal, but that LASSO is less affected by the ablation procedure compared to Random Forest.

5.4.3 Crash Indicator Mechanistic Interpretability



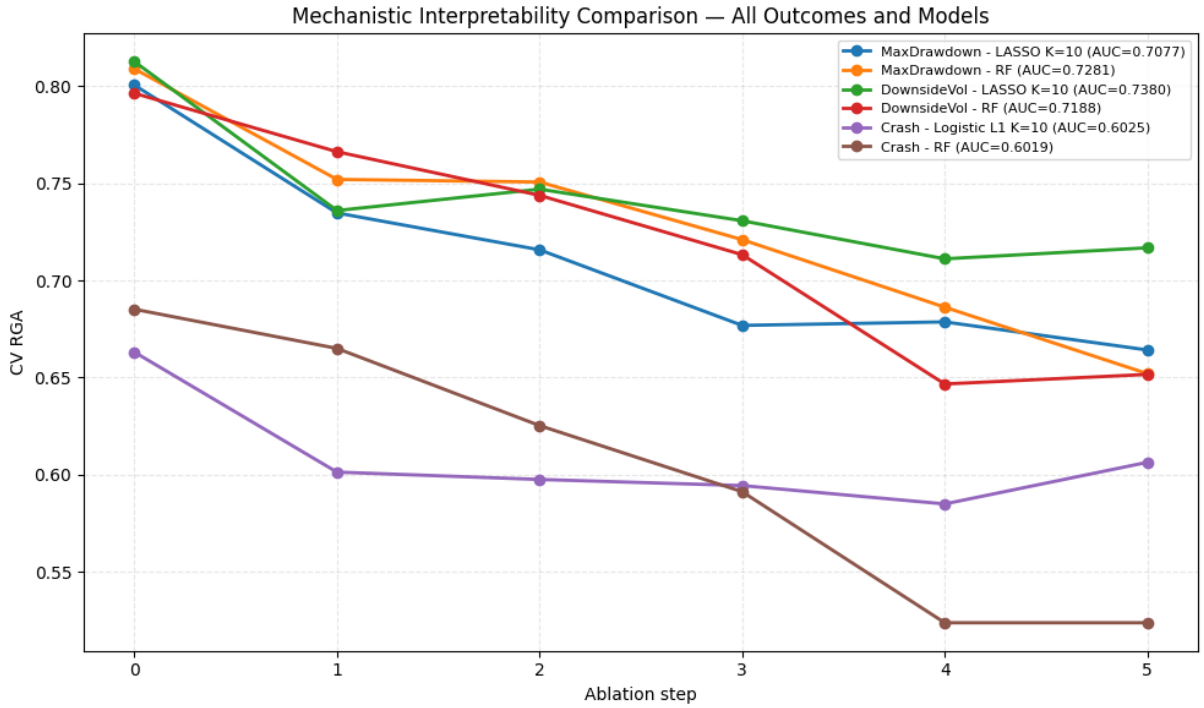
Since for the crash indicator a classification is required, the model used is Logistic L1. Here the baseline cross validated RGA is 0.6631. After removing volatility, the RGA drops to 0.6013. Then the model removes the Financials sector and the RGA becomes 0.5975. After that, momentum is removed and the RGA becomes 0.5943. Subsequently, the model removes the Health Care sector and the RGA falls to 0.5849. Lastly, the Industrials sector is removed and the final RGA increases to 0.6063. The total drop in RGA from the baseline model is 0.0568. The AURGAE is equal to 0.6025. It is noticeable that the removal of the last variable temporarily improves the RGA. Overall, this shows that the crash ranking is more irregular and that some variables may not contribute consistently to the ranking.

For the Random Forest, the baseline cross validated RGA is 0.6852. After removing volatility, it drops to 0.6650. Subsequently, momentum is removed which brings the RGA to 0.6252. Next, the logarithm of the market capitalization is removed by the model and brings the RGA to 0.5912. Afterwards, the market beta is removed which brings the model to an RGA of 0.5237. The last and fifth variable removed by the model is the ESG score, where the RGA falls to 0.5237. The total drop from the baseline RGA to after the removal of the five variables corresponds to a total fall of 0.1815. Here the AURGAE is 0.6019. It is observable, that the Random Forest starts at a higher RGA level than Logistic L1. Additionally, the decline of the Random Forest model becomes very large in the later removals, so that its final RGA ends below that of Logistic L1. Empirically, Random Forest is more sensitive to the cumulative removal and it depends more strongly on information that surfaces by combining variables.

Comparing the two models, the baseline RGA is higher for Random Forest, while the final RGA is higher for the Logistic L1 model. The total drop from the baseline RGA to the RGA after the five variables are removed is 0.0568 for Logistic L1 and 0.1615 for Random Forest. This contrast shows how Random Forest loses substantially more RGA than Logistic L1. The crash indicator is the variable that shows the strongest difference between the baseline advantage and the ablation

robustness. The empirical conclusion is that Random Forest ranks crash better initially, but that Logistic L1 is less sensitive to information removal and doesn't drop as strongly.

5.4.4 Overall Comparison Across Outcomes



The general pattern for the crash indicator is that the feature removal reduces the cross validated RGA for both Random Forest and LASSO/ Logistic L1. All outcomes show dependence on the selected predictors. The curves confirm that the performance of the model is not independent of the information that the variables add. Overall, the ablation analysis provides additional inside beyond the simple baseline RGA calculation.

The Random Forest has the larger total drop in RGA score for all three outcomes, while LASSO and Logistic L1 retain more RGA even after removing variables. This shows that Random Forest is more sensitive to cumulative ablation, while LASSO appears more stable when removing features.

On the outcome level, the LASSO model for the downside volatility has the highest baseline RGA. For maximum drawdown, both models lose substantial performance. Both have a similar RGA at the baseline but differentiate after feature removal. Lastly, the crash indicator has the most irregular ablation curve, not just decreasing but also increasing after some variable removals. Logistic L1 shows a non-monotonic pattern, even though the final RGA after ablation is just below the initial baseline RGA. Random Forest, on the other hand, loses the most total RGA.

Overall, the mechanistic interpretability analysis shows that removing important variables and predictors reduces the performance across all models. Random Forest generally shows a larger RGA drop than LASSO and Logistic L1, implying a higher sensitivity to feature removal. Contrarily, LASSO retains more ranking performance even after ablation, making it a more reliable and interpretable model for downside risk ranking.

5.5 A Multidimensional Downside Risk Framework

5.5.1 Integrated Model Comparison

The prior sections in chapter 5 were about evaluating the models in four separate dimensions: the baseline ranking performance (RGA), the subset robustness (AURGA), the noise robustness (AURGAR) and the mechanistic interpretability (AURGAE). Till now the metrics were looked at individually, which makes it difficult to determine which model performs best overall. Therefore, an integrated comparison is conducted to unify the four dimensions into one single measure.

To achieve this, four means are chosen: the arithmetic mean, the geometric mean, the harmonic mean and the quadratic mean (RMS).

The arithmetic mean is it used to trade all four dimensions equally. It serves as an intuitive and simple benchmark to measure the overall performance. The formula used for this is

$$AM = \frac{RGA + AURGA_{Subset} + AURGAR_{Noise} + AURGAE_{Interp}}{4}$$

The purpose of the geometric mean is to penalise more the dimensions that are weaker. This way, a balanced performance is achieved. Using the geometric mean it is made sure that the model cannot compensate for a poor performance by excelling somewhere else. The following formula was used.

$$GM = (RGA * AURGA_{Subset} * AURGAR_{Noise} * AURGAE_{Interp})^{\frac{1}{4}}$$

The third mean used is the harmonic mean to penalise more models that are performing poorly. Through this, weaknesses are highlighted. The formula is

$$HM = \frac{4}{\frac{1}{RGA} + \frac{1}{AURGA_{Subset}} + \frac{1}{AURGAR_{Noise}} + \frac{1}{AURGAE_{Interp}}}$$

The last mean used is the quadratic mean (RMS). Its objective is to reward strong performances, By giving more weight to high values. Its formula is

$$QM = \sqrt{\frac{RGA^2 + AURGA_{Subset}^2 + AURGAR_{Noise}^2 + AURGAE_{Interp}^2}{4}}$$

These four measures are useful to provide robustness to the evaluation. If the ranking of the models remains the same even across the arithmetic, geometric, harmonic, and quadratic mean, that signifies that the end result is unlikely to depend on a particular method used. It makes sure, that the outcome is firm and resilient.

The results from this analysis are summarised in the subsequent table:

Outcome	Model	Arithmetic	Geometric	Harmonic	Quadratic
MaxDrawdown	LASSO K=10	0.7234	0.7186	0.7137	0.7279
MaxDrawdown	Random Forest	0.7310	0.7254	0.7194	0.7362
DownsideVol	LASSO K=10	0.7394	0.7345	0.7294	0.7439
DownsideVol	Random Forest	0.7183	0.7124	0.7061	0.7238
Crash	Logistic L1 K=10	0.6032	0.5988	0.5940	0.6074
Crash	Random Forest	0.6146	0.6091	0.6033	0.6197

These outcomes are categorised by a downside measure and for each mean the winning machine learning model is chosen. This is depicted in the following table:

Outcome	Arithmetic_Mean	Geometric_Mean	Harmonic_Mean	Quadratic_Mean
MaxDrawdown	Random Forest	Random Forest	Random Forest	Random Forest
DownsideVol	LASSO K=10	LASSO K=10	LASSO K=10	LASSO K=10
Crash	Random Forest	Random Forest	Random Forest	Random Forest

From this table one clear winner per downside model can be chosen. The choice is straightforward since for every downside of metric all four means have the same winning model.

Outcome	Winner
MaxDrawdown	Random Forest
DownsideVol	LASSO K=10
Crash	Random Forest

This table shows that for both maximum drawdown and the crash indicator, the winning model is Random Forest. It has the highest means. On the opposite, for downside volatility LASSO K=10 has the advantage.

Analysing the result more individually, the arithmetic mean is the one chosen because of its simplicity and straightforwardness. Additionally, it is the one that has overall the highest scores for all three downside measures.

Maximum drawdown has a score for Random Forest off 0.7310 and for LASSO of 0.7234. This result explains that Random Forest is able to consistently achieve a higher overall performance, even if just by slightly since the advantage is moderate. Overall, both models perform strongly, indicating that they are both adequate to predict downside risk.

For downside volatility LASSO has an arithmetic mean of 0.7394, while Random Forest has it of 0.7183. This is at the largest difference between all three outcomes. This shows that LASSO offers a better performance and that the linear model dominates over the nonlinear one.

The third and last metric is the crash indicator, where the arithmetic mean of Random Forest corresponds to 0.6146 compared to LASSOs 0.6032. Like for the maximum drawdown, the Random Forest model wins here as well. It has to be noted though, that both scores are lower than for the continuous measures, indicating that the models have more difficulties in predicting binary crash events.

This comparison reveals that no model dominates all three downsides of risk measures. Random Forest performs better for maximum drawdown in crash indicator, while LASSO wins when analysing the downside volatility. It is important to note though, that the model rankings remain identical for all four different means. This highlights that the results are robust and spread.

5.5.2 Comparison across Downside Risk Measures

While the prior section identified the best machine learning model for each downside risk measure, it didn't evaluate the overall best predicted downside risk. To address this, the composite scores of the performance of the single models were compared across maximum drawdown, downside volatility and the crash indicator.

The following table presents the three overall winning scores for each measure. Here as well, they are compared based on their arithmetic mean.

Outcome	Best Model	Overall Score
DownsideVol	LASSO K=10	0.7394
MaxDrawdown	Random Forest	0.7310
Crash	Random Forest	0.6146

Looking at the results more in depth, it is observable that downside volatility is the model with the highest overall scores of 0.7394. This shows that downside volatility has a strong baseline RGA, as well as strong robustness result. Additionally, it has as stable interpretable performance and it performs consistently well across all dimensions that are evaluated.

Maximum drawdown is the second scoring model with 0.7310. It is just slightly behind the downside volatility. A present as well a strong predictive performance, high robustness and a good overall score. The difference between downside volatility and maximum drawdown is only off 0.0084, which is economically negligible.

The last score is the crash indicator, which show is a much lower overall result compared to the two continues measures. It indicates a lower baseline RGA and a lower robustness, as well as a greater sensitivity to changes in the model. This is an expected result, since crashes are rare events and a binary variable, which makes it more difficult to evaluate and may not be fully reflected in firm characteristics.

Concluding, the overall ranking is downside volatility, then maximum drawdown and at last the crash indicator. The numbers show that the continuous measures are better predicted by the models end that the firm characteristics contain stronger predictive information. On the other hand, the binary risk indicator is less effective and more difficult to forecast.

5.6 Key Empirical Findings

Summing up the whole chapter 5, in sub chapters 5.1 to 5.4 the models were evaluated from different angles. They were analysed through the baseline ranking performance, the subset robustness, the noise robustness and the mechanistic interpretability. Then in 5.5 the models were compared through four different means, resulting in a winning model for each downside risk measure and lastly a classification of the easiest to predict downside risk.

The first finding is that downside risk is predictable. All models achieve a meaningful RGA value, Especially maximum drawdown and downside volatility have a very strong baseline RGA values. This indicates that the firm level characteristics from 2023 contain good predictive information to rank 2024 downside risk.

The second finding is that the continuous downside measures are easier to rank than the crash indicator. This comes out very clearly in the final model comparison in 5.5.2, Where downside volatility has a strongest overall score of 0.7394, maximum drawdown is a close second with 0.7310 and the crash indicator has a much lower score of 0.6146. These results show that the continuous variables are easier to predict and contain more information. Crash events are rare and harder to predict and the fact that it is a binary variable adds to the difficulty of prediction.

The third finding is that there is not one single machine learning model that dominates every result. Maximum drawdown and the crash indicator are best predicted by the Random Forest, while the downside volatility achieves the highest score through LASSO K=10. It can be deducted that the model choice depends on the downside risk measure and there is not one superior model for all three downside risks.

The fourth finding is that Random Forest is a flexible model, but that LASSO is more transparent. Random Forest Is able to capture nonlinearities and interactions, while LASSO is sparser and more interpretable. Overall, it is often noticeable that LASSO maintains a stronger performance even though it is a linear and simpler model.

The fifth finding is that when removing variables from the model, the predictability score decreases. When only a few variables are removed the model normally still performs at a high level, but when more and more variables are delated from the sample, the model finds it harder to predict future downside risk. This confirms that some firm characteristics are strong predictors for the future downside risk.

The sixth finding is that the robustness results support the baseline conclusions. The subset removal analysis showed that when more predictable firms were removed from the sample, the performance of the model decreased. Moreover, it depicted that the crash risk is the most fragile, while continuous downside measures remain more stable.

Concluding, the empirical results show that firm characteristics contain important information to predict future downside risk. The evidence is stronger for the continuous measures maximum drawdown and downside volatility, while for the binary crash indicator the prediction is harder. Furthermore, Random Forest provides the best overall performance for maximum drawdown and crash indicator, whereas LASSO performs best for downside volatility. This supports the statement that the model choice should be made based on the downside risk objective and measure. Lastly, the analysis of interpretability and robustness are valuable evaluations to add to the baseline performance ranking.

6. Economic Discussion

6.1 Introduction

In the traditional finance world, the average returns are normally the main focus. Investors though are often more concerned about potential losses than possible gains, since large negative returns can impact portfolios and have important consequences. Therefore, downside risk is a significant metric and analysis for investors, portfolio managers, risk managers and regulators.

The idea is that investors don't experience risk symmetrically. The same percentage incurred in return or loss does not have the same economic impact. For this reason, the maximum drawdown, the downside volatility and the crash risk are important measures for the finance and investment world.

In the empirical result analysis from Chapter 5 it is shown that the downside risk can be predicted through machine learning models, since firm level characteristics contain information about the future downside threat. The prediction quality though defers depending on the model and on the downside variable measured. Overall, the downside volatility and maximum drawdown achieve higher predictability scores than the crash indicator.

For now, the results have been only statistically interpreted and evaluated. This does not imply though, that they are economically significant and it doesn't explain why these relationships exist. For this reason, it is essential to understand the economic reasoning behind these results.

The first objective of this chapter is to explain why the downside risk is predictable. Secondly, the differences between the three downside risk metrics are interpreted. The third objective is to understand and discuss the economic role of the most important predictors from the ablation analysis, especially the volatility, the market beta, the momentum, the market capitalization and the sectors. Lastly, the implications for investors, portfolio construction and risk management are discussed.

6.2 Why Is Downside Risk Predictable?

The predictability downside risk is an important topic, especially since the prediction of future risk should not be trivial.

6.2.1 Persistence of Firm Characteristics

The model uses firm characteristics to predict future downside risk. These from characteristics are the market beta, the volatility, the momentum, the market capitalization, ESG variables and information about the size of the company. These variables don't randomly change every year, they tend to persist or adapt to the market conditions. For this reason, a company that in the present is highly volatile, very leveraged and operationally weak, will tend to maintain these characteristics one year from now as well. Subsequently, if a company is risky today it will be very likely still risky one year later. Therefore, the 2023 characteristics contain information about the 2024 downside risk.

This explanation is consistent with the fact that downside volatility and maximum drawdown were both highly predictable. This means that the firm risk can't be completely random, otherwise the models would not have been able to achieve these high results.

6.2.2 Persistence of Volatility and Risk

In finance and economics, volatility has always been one of the most important variables, since it influences risk, expected return and pricing. Volatility is a variable that is known to cluster, meaning that periods of high volatility are normally followed by periods with high volatility as well. This happens because high volatility reflects uncertainty, disagreement between investors, uncertain outcomes and changing conditions. These fundamentals rarely adjust immediately.

These findings are supported by this thesis since volatility repeatedly emerges as the dominant predictor. By removing the volatility in the mechanistic interpretability section, the model suffered from a decline in performance and a strong decrease in RGA. The preliminary analyses that consistently identified volatility as the most influential predictor, motivated the targeted removal in the ablation analysis. By eliminating volatility first, the effect of keeping all the other variables besides volatility came to the front. Hence, the importance that emerged from the volatility variable in the empirical analysis is consistent with the persistence of risk that is present in financial markets.

6.2.3 Continuous Risk is Easier to Forecast than Rare Events

Furthermore, continuous risk is easier to forecast than rare events. This is observable in chapter 5, where the overall scores for each model are 0.7394 for downside volatility, 0.7310 for maximum drawdown and 0.6146 for the crash indicator. This result clearly shows that the continuous downside measures are simpler to predict for the model.

Continuous measures evolve gradually and reflect the underlying conditions off the firms. On top of this, they accumulate over time. On the other hand, crash events are rare, they occur suddenly and are often triggered by unexpected information. For example, crashes can arise from accounting scandals, regulatory news, macroeconomic shocks or earnings surprises. These factors are hard to anticipate in advance, especially long periods of time like a year can be.

6.2.4 Informational Content of Firm Characteristics

The analyses in this thesis show that the predictors chosen contain economically meaningful information. The market beta for example measures the market sensitivity. A higher beta means that the market reacts stronger to downturns. If the market capitalization is small, it indicates that the firm has a lower diversification, a greater vulnerability and more constraints. The momentum shows that past poor performance worsens the fundamentals and makes investors more pessimistic. ESG variables can capture the governance quality, the operational capacity of a firm and how effective management is.

These firm characteristic variables are not just statistical inputs, they actually reflect the underlying economic conditions of the market and the company, as well as the financial resilience and the quality of the firm. Therefore, it is economically reasonable that these variables help in predicting future downside outcomes.

6.3 Economic Interpretation of the Three Downside-Risk Measures

In Chapter 5, the empirical results of the analyses where presented, which showed that the model differs in predicting future downside risk. Downside volatility was the measure that achieved the overall highest score and maximum drawdown was a close second. Conversely, the crash indicator score was considerably lower. These differences suggest that these measures capture different dimensions of downside risk and use more or less the predictable information.

6.3.1 Downside Volatility

Downside volatility achieves an overall score of 0.7394, which is the highest score between the three models. The reason for why downside volatility is so predictable is because of what it captures economically. Downside volatility seizes the variability of negative returns, the frequency of these fluctuations and the instability during bad periods. It is important to mention that it is not a single event, but it is an ongoing process. This is why it is considered as a continuous measure.

The first reason for why downside volatility is predictable is that it shows a gradual deterioration. The risk normally doesn't happen all at once, it usually accumulates; for example, when a company's profits decline, when the cash flow becomes weak, when the uncertainty increases or when business conditions become worse. These are all situations where change happens gradually. As consequence, the market incorporates these changes progressively. Therefore, future downside volatility is partially already visible in today's characteristic.

Second, uncertainty is persistent, which means that a high today will likely cause a high uncertainty tomorrow as well. For this reason, in periods with high uncertainty investors remain more cautious, demand a higher risk premium and react stronger to bad news.

Third, the results in the mechanistic interpretability section show that when volatility is removed, the model deteriorates noticeably. Subsequently, volatility is a good indicator since it often signals weak expectations, uncertainty in cash flows and an increased downside exposure. Thus, including volatility in the predictions makes it easier for the model to forecast future downside volatility.

For investors this signals that downside volatility is not a random measure. Vulnerable firms can be identified early and the management of risk can be useful. Additionally, by predicting future downside volatility, investors can adjust their portfolios in advance, they can control the risk and monitor endangered firms more closely.

Since downside volatility is highly predictable, it suggests that it is driven by firm characteristics and therefore represents a forecastable measure of downside risk.

6.3.2 Maximum Drawdown

From the empirical results section the overall score for maximum drawdown for 0.7310. This is a very close number to the one of downside volatility, the difference is only of 0.0084.

Maximum drawdown measures the largest cumulative loss that an investor would incur if he held a specific stock in that specific period of time. It is the worst peak-to-trough loss and it shows the

severity of bad market episodes. Contrary to volatility, it focuses on extremes and more severe losses. From the results, maximum drawdown remains highly predictable.

The first reason for why maximum drawdown is highly predictable is that vulnerability accumulates over time. Important drawdowns rarely come out of nowhere, they have been building up from past periods. Drawdowns often reflect worsening fundamentals of firms, very excessive risk taking or uncertainty that spreads over time. All these characteristics are noticeable beforehand.

Secondly, drawdowns are continuous. Even though the maximum drawdown is a concrete extreme point in time, a company can lose already at 20%, 40% or 60%. This indicates that the model maintains knowledge. For this reason, a continuous model outcome Preserves more information than a binary outcome.

Third, firms are often fragile and drawdowns often reveal information about the firm's vulnerability. They bring to the surface if a company has a weak business model, is constrained financially or has operational weaknesses. These points brought to the surface by drawdowns are exactly the characteristics captured in this thesis' predictors.

Investors care about capital preservation and they want to avoid large cumulative losses. Especially for long term portfolios, short term volatility is put to the side and long term performance is the focus. Therefore, anticipating maximum drawdown risk adds practical value to investors and risk managers. The high predictability of this measure highlights that severe cumulative losses are not the result of random events or shocks, but that they are closely linked firm characteristics and indicators of companies' performance.

6.3.3 Crash Indicator

From the empirical results the overall score of the crash indicator is 0.6146. This is a substantially lower result than both of the continues outcomes had achieved, the binary result is over 0.12 below the other two measurements.

The crash indicator is built different from the two continuous measures. It is binary, discrete and based on events, since a firm either crashes or doesn't crash. For this reason, a lot of information disappears from the variable and few predictors are kept. It is the deductible, that the crash indicator is much harder to forecast than the downside volatility or maximum drawdown. This is due to a few economic reasons.

The first reason is that crash events are rare episodes. They occur infrequently, for example when macroeconomic shocks happen, when speculative bubbles burst or when monetary policies change. Therefore, the model has fewer observations and less information to base its predictions on. This reduces the predictability of the algorithm.

Second, information is often unexpected. Crash triggers can be earning surprises, fraud revelations, regulatory actions or litigations. These events are often unforeseeable and can disrupt the economic equilibrium.

The third reason is the timing problem. Firm characteristics can frequently reveal whether a company has vulnerabilities, is it has a higher probability to be more volatile in the future or incur more downside risk. They cannot reveal though whether or when a shock will exactly arrive in the future. An example for this is that a company that is bad on paper can survive for years even

though it is weak financially, doesn't have a good management or is not well positioned in the market. On the other hand, a company that is strong and well established can crash after an unexpected event. Thus, since the timing of the crash is difficult to forecast, the whole measure becomes more unsure and more difficult for the model to predict.

Mainly, most of the crash risk comes from new information, but new information cannot be perfectly predicted. For this reason, investors are often able to identify vulnerable and risky firms, but they cannot perfectly predict crash events.

The lower predictability of the crash indicator suggests that firm characteristics are primarily adequate to capture vulnerability and distress, but that they have more difficulties in predicting extreme negative shocks.

These results align with financial theories like market efficiency and risk literature.

Summing up, downside volatility is the most predictable measure in maximum drawdown is a very close second. The crash indicator on the other hand is less predictable and has a lower model score. This difference arises because the two continuous variables evolve gradually and with time, while the crash event depends on unexpected information.

6.4 Economic Interpretation of the Main Predictors of Downside Risk

In section 6.3, the focus was on the three downside risk measures. This section shifts the attention to the explanatory variables. In the mechanistic interpretability analysis and the feature important evaluation the results identified multiple variables that contribute to the performance of predictions. To get an accurate explanation an interpretation, it is important to understand why these variables matter and their influence on the forecasting of downside risk measures.

6.4.1 Volatility

In the mechanistic interpretability section, the ablation analysis was structured to evaluate the sensitivity of the prediction performance of the model, when firm characteristics and economic variables are removed. The code was structured to remove Volatility 2023 intentionally first. This was done because preliminary coefficient and feature importance analyses have found out, that volatility is a consistent dominant driver and predictor of future downside risk. By removing volatility first, it is observable how the model reacts to not being able to rely on an important predictor, even though all the other variables are still present.

In Chapter 5, emerges as one of the strongest predictor when observing the decrease in RGA score after the removal. It's elimination from the model causes a noticeable drop in performance. This is observable for all downside risk measures.

Volatility matters as relevant economic variable since it stands for numerous information. It reflects the uncertainty about future cash flows, disagreement between investors, changing expectations and economic environment, as well as unstable business conditions.

Economically, whenever high volatility is present, it signals that investors are uncertain, that the information is possibly noisy and that future outcomes are difficult to predict due to changing situations. For these reasons, firms that have a high volatility are more prone to downside events.

From financial theories, it is known that volatility clusters, meaning that periods of high volatility are followed by other periods of still high volatility. This can be often found as well in financial markets, where there is persistence in volatility and clustering of uncertainty. Subsequently, past volatility contains information about future downside risks.

The result of this thesis are coherent with economic theories. Models show a strong predictive performance and a strong ablation effect, especially the continuous outcomes.

6.4.2 Market Beta

The market beta consistently shows up in the results, meaning that it contributes to predictions.

The beta measures the sensitivity of a company to market movements, so a firm that has a high beta reacts more strongly during bad times and experiences larger losses in distress. For this reason, a higher beta implies higher downside exposure.

This can be proven through the capital asset pricing model (CAPM) and systematic risk, where a firm that is highly exposed to market risk is more affected by a market decline and economic downturns.

6.4.3 Firm Size (Market Capitalization)

Another variable interpreted as one of the driving predictors is the market capitalization. It is used as proxy for the firm size, since it represents the total market value of a company's outstanding shares. The higher the market capitalization of a company is, the larger is its position in the market.

The economic interpretation behind it is, that the bigger a firm is, the more it benefits from higher diversification, easier access to financing and stronger liquidity. On the other hand, if a company is smaller it often faces financing constraints, fragility in operations and more uncertainty. As a consequence, smaller firms can experience bigger drawdowns, a higher volatility and more vulnerability during market distress. For these reason, the size of a company is often an indicator for how resilient and well established the firm is.

6.4.4 Momentum

The momentum appears as well among the important predictors of downside risk. It is the tendency of the price of an asset or the performance of a company to continue moving in the current direction.

The momentum of a company reflects the recent market sentiment towards it, how investors perceive it and the continuation of trends. When the momentum of a firm is substandard, it typically indicates that the fundamentals of the company are deteriorating and that the expectations of investors are negative. As a consequence, this increases the downside vulnerability and risk.

The importance of momentum is often explained in behavioural finance, which helps to explain why past and present information are relevant for future predictions.

6.4.5 ESG Variables

ESG related variables are included as firm predictors, but they aren't between the dominant forecasting measures. They appear only twice in the five-step ablation analysis.

In economics, ESG variables represent governance quality, environmental resilience and social factors. If the governance quality is high for a firm, it shows that they have a strong monitoring, high risk controls and few agency problems. When firms have a low environmental resilience, they are more prone to face litigations, regulatory issues and damage their reputation. The social performance on the other hand can affect their customer relationships, how they retain their employees and how stable their operations are.

In the analysis it becomes clear that ESG firm characteristics can contain information about the quality of the organisation and how resilient it is, but that they are less relevant for downside risk forecasting. The typical financial factors prevail over ESG variables in forecasting future downside exposure.

6.4.6 Sector Effects

Several sector dummies are removed in the ablation analysis, meaning that the model sees them among the top five predictors. Especially the financials, healthcare, infotech, real estate, industrials and utilities sector are between They remove variables.

The membership of a firm to a specific industry can influence its downside risk because different sectors have different regulatory environments, business cycles and exposure to macroeconomic events.

An important limitation to mention is however that sector dummies act primarily as control variables rather than actual risk measurement variables. For this reason, their importance should be interpreted as variables that can capture systematic differences between industries instead of actual firm specific characteristics.

6.4.7 Conclusion

Overall, it is important to note that all these predictors share a common aspect. They all capture uncertainty, vulnerability and resilience of firms.

Because of this, the models don't present random statistical patterns, they actually fined meaningful signals for forecasting downside risks.

6.5 What do the Model Differences Mean?

The purpose of this section is to explain why Random Forest wins for both maximum drawdown and the crash indicator and why LASSO prevails for downside volatility. These results were obtained at the end of chapter 5. Overall, no model dominates all three outcomes. This indicates that the most successful model depends on the type of downside risk. From this it can be deducted that the three downside risk measures capture different economic mechanisms.

Random Forest performs better for maximum drawdown mainly because this downside risk is path dependent. It captures the worst peak-to-trough loss. Drawdowns often depend on a multitude of variables, for example on volatility, market beta, firm size or momentum. From the results it is deductible that Random Forest is able to capture these interactions better than LASSO, since it is nonlinear.

LASSO prevails for downside volatility, which is the downside measure with the strongest overall score. This indicates that the measure is driven by a stable and compact group of predictors. LASSO is useful here since it is able to select a sparse group of variables. It is a linear model and can capture regular and foreseeable risk.

Lastly, Random Forest performs better for the crash indicator, which is a binary variable and rare event. In this thesis it is defined as the worst 10% annual returns, which corresponds to 42 firms of the sample. Crash events can depend on thresholds and non-predictable events, for which a nonlinear model works best.

Overall, while LASSO is easier to interpret and selects fewer variables, Random Forest is more flexible and can capture nonlinear interactions, even though it is harder to interpret.

Since there is not one machine learning model that is superior, the model's choice should depend on the economic objective and variable. LASSO should be used when the interpretability is crucial, while Random Forest should be chosen when nonlinear interactions have to be captured.

6.6 Implications for Investors and Risk Managers

The purpose of the section is to explain why it matters, that the models in the thesis are able to predict downside risk. In practice, investors don't need the exact prediction of losses, often estimations and directions are enough to establish a stable financial investment position. Additionally, it is not useful to simply identify which firms are more vulnerable than others. For this reason, these models can be used as a screening tool and the firm's results to have a higher predicted downside risk can be given closer attention. This is a useful structure for portfolio monitoring, risk analysis and asset allocation. The key idea is it that the economic value added by these models is not to perfectly forecast numbers, but to improve the prioritisation between companies.

As the results from Chapter 5 indicate, the continuous risk measures are better at assessing downside risk from the early stages on. Recapitulating the results, downside volatility has the strongest overall score with maximum drawdown being a very close second. The crash indicator is significantly weaker.

From this it can be taken that downside volatility and maximum drawdown are better measures for proactive management and monitoring. The crash indicator is still a Use for signal, but harder to forecast. Because of this, it is best for risk managers to not rely simply on the binary crash predictions.

From the ablation analysis it is clear that volatility is a dominant risk variable, since removing it from the firm characteristics decreases the performance significantly. This confirms that past volatility contains important signals for future downside risk forecasting. Subsequently, investors should monitor firms with an elevated prior year volatility first, since they're high volatility can indicate instability, uncertainty and vulnerability.

Moreover, depending on the goal of that usage, different machine learning models should be applied. If the goal is for example transparency, LASSO should be the primary tool, while if the goal is a flexible prediction, then Random Forest should be used. Therefore, investors should choose the model according to their decision problem.

The inclusion of sector dummies in the analysis is important, since sector controls are important for risk monitoring. They account for the fact that different sectors have a different structure and perceive different risks. Without them, the model could attribute sector effects to firm variables. Following this, the risk should not be evaluated only through firm characteristic but industry exposure should be included as well. For example, a firm that has a high risk in a sector which is also highly unstable, requires additional monitoring.

Furthermore, ESG variables can complement financial variables, since they contain information that goes beyond the traditional signals. It can add predictive power by capturing the governance quality, the resilience of a firm and the exposure to non-financial risk. It is important to note though, that the main financial variables are still the most important predictors. ESG variables should only complement and enrich the downside risk evaluation, not completely replace the traditional financial risk measures.

This analysis does not propose a perfect prediction system. It proposes a decision tool based on ranking. The value of the model comes from it being able to identify early firms that are going to be vulnerable in the future and as a consequence have a higher downside risk. This is very useful for predictions in the economic field, since financial markets are noisy and extracting forecasts is difficult.

7. A Multidimensional Risk Ranking Framework

7.1 Introduction

The focus in chapter 5 and chapter 6 is more on the evaluation of the models and on their economic interpretation. Their objective is to determine whether it downside risk is predictable, which models perform best and which room characteristics actually drive downside risk. Ultimately, investors require a practical tool to apply these findings and use them for their decision-making process. For this reason, a new framework is needed.

Downside risk is multi-dimensional and as seen in the previous chapters, there is not one single downside risk measure that can fully capture the whole aspect of risk. While the maximum drawdown captures the severity of potential losses, downside volatility portrays the frequency and magnitude of negative fluctuations. On the other hand, the crash indicator sizes how likely extreme downside events are. Consequently, investors are interested in all three dimensions simultaneously, since a firm can exhibit multiple downside risks at once.

Due to the importance of all three downside risk measures, the purpose of this chapter is to combine them into one metric. This is done through the combination of the strongest predictions from each downside risk dimension into a single multi-dimensional ranking framework. The strongest predictions for each downside model were analysed in chapter 5, where for maximum drawdown and the crash indicator Random Forest was selected, while for downside volatility LASSO dominated. The selected models are now carried forward in the decision framework, no additional model selection is performed in this chapter.

The framework is designed to transform the predictive outputs into risk assessments that allow investors and managers to take actions. For this reason, the sole objective is not only the prediction, but also the support of investment decisions. Here the focus shifts from a model evaluation perspective to a practical application.

7.2 Construction of the Multidimensional Risk Score

The three prediction outcomes are measured on at different scales. Therefore, a direct comparison is impossible, because drawdown is a percentage loss, downside volatility is a volatility measure and the crash indicator is a binary probability. Subsequently, normalisation is required before aggregating the three measures. The aggregating method chosen is the percentile transformation, which still preserves all the information relative to the ranking.

To make these three comparable, their predictions are transformed into percentile ranks. If for example a firm is in the 95th percentile, this shows that it is riskier than 95% of the companies in the sample. Through this comparison method, a common scale ranging from 0 to 100 is created.

Once the measures are converted into percentiles, the three dimensions can be combined. Through this ensemble aggregation, the framework integrates the prediction of drawdown risk, downside volatility risk end crash risk into one overall downside risk score.

It is important to mention that different investors may have different importance preferences. Some may want to put more weight on drawdown, volatility or crash. To accommodate this, investors can choose multiple weighting schemes to place more or less importance on the individual downside risk dimensions.

The general formula is

$$Risk\ Score_i^{Mean} = w_1 * Pct\ MaxDrawdown_i + w_2 * Pct\ DownsideVol_i + w_3 * Pct\ Crash_i$$

which is subject to

$$w_1 + w_2 + w_3 = 1$$

To avoid the introduction of subjective preferences, this thesis adopts equal weights. Thus, for the empirical analysis it holds that

$$w_1 = w_2 = w_3 = \frac{1}{3}$$

In this analysis, equal weights are chosen for neutrality. This avoids a subjective prioritisation of a one of the three downside risk dimensions. Here, all three measures are treated as equally important. An alternative weight structure is possible depending on the objective and preferences of the single investors.

Because the crash indicator risk is based on a binary outcome, it is possible that the extreme values affect the averages. Therefore, both the arithmetic mean aggregation and median aggregation are used. Through this it can be assessed whether the firm rankings remain stable, even with alternative aggregation methods.

7.3 Identification of High-Risk Firms

This multidimensional risk analysis is applied to the S&P 500 companies, to achieve a firm level ranking. Through this, the firms with the highest predicted downside risk are identified based on the aggregate risk score.

The equal weighted arithmetic mean is used as the main tool for the ranking, since it incorporates information from all three dimensions without ignoring any downside risk measure. The median ranking is computed as well to check the robustness of the results.

It is important to mention, that the relative risk is a ranking within the sample not an absolute probability of loss or a risk evaluation on the whole market.

The following table shows the top 10 firms by mean ranking. The firm that is in rank one is the company with the highest predicted multi-dimensional downside risk. Its prediction is based on the three percentile dimensions, namely the forecast of the maximum drawdown risk, the downside volatility risk and the crash indicator risk. It is an investor facing output to identify the company is that require closer monitoring.

Ticker	Company	Sector	Pct_MaxDrawdown	Pct_DownsideVol	Pct_Crash	Ensemble_Risk_Mean	Final_Risk_Rank_Mean
ALB	Albemarle Corporation	Materials	99.76	98.31	99.76	99.28	1
NCLH	Norwegian Cruise Line Holdings	Consumer Discretionary	99.28	98.55	94.20	97.34	2
IFF	International Flavors & Fragrances	Materials	98.31	93.96	99.52	97.26	3
EL	Estée Lauder Companies (The)	Consumer Staples	96.86	96.62	96.14	96.54	4
MOS	Mosaic Company (The)	Materials	99.03	88.65	98.79	95.49	5
APA	APA Corporation	Energy	93.72	93.24	99.28	95.41	6
BBWI	Bath & Body Works, Inc.	Consumer Discretionary	91.55	95.89	98.07	95.17	7
FMC	FMC Corporation	Materials	100.00	84.78	100.00	94.93	8
STLD	Steel Dynamics	Materials	97.34	87.68	99.03	94.69	9
ON	ON Semiconductor	Information Technology	97.10	99.76	86.71	94.52	10

Through aggregate scores, the risk ranking and percentile values are based on several downside risk dimensions. Consequently, the analysis is not done only through one isolated metric. The

companies that are ranked with high risk is because of their combined vulnerability to the three scores. This way, a multidimensional downside exposure ranking is created.

According to the proposed framework, Albemarle Corporation (ALB) has the highest ensemble downside risk score, since it shows very high percentiles in all three dimensions. It has a strong multi-dimensional risk profile.

Evaluating the top 10 riskiest companies in the sample by mean ranking, shows that many high risk firms are from cyclical or commodity sensitive sectors, like materials, consumer discretionary, energy and information technology. This can be due to the fact that firms in these industries are exposed to changing input prices, cycles in macro shocks, as well as demand changes and evolving expectations. As mentioned in the prior chapters, the sector exposure of companies can potentially amplify the firm level downside risk.

7.4 Robustness: Mean vs Median Aggregation

As priorly mentioned, both the mean and median aggregation are computed to check robustness and whether the final firm ranking depends on the aggregation method. For this, the arithmetic mean ranking is compared with the median ranking, to test the sensitivity of the model to extreme percentile values. Through this it is verified whether the high-risk firms remain similar even with different aggregation.

In this analysis the median is useful since it is less affected by an extreme percentile value. This is relevant because the crash probability comes from a binary outcome. Through the median, the ranking is protected from a very high or very low risk dimension, since it focuses on the typical risk level across the three dimensions. No weighting scheme is required. Overall, it is a useful alternative to the arithmetic mean. Its formula is the following one

$$Risk\ Score_i^{Median} = Median(Pct\ MaxDrawdown_i, Pct\ ownsideVol_i, Pct\ CrashIndicator_i)$$

The mean is still used as the main ranking for this evaluation since it uses information from all three dimensions, making sure that all percentile scores contribute directly. It is a better tool for capturing the total multidimensional exposure. The median is rather used as a validation instead of a replacement.

This thesis compares the top 10 firms ranked by the ensemble risk mean and the top 10 firms ranked by the ensemble risk median.

Ticker	Company	Final_Risk_Rank_Mean	Final_Risk_Rank_Median
ALB	Albemarle Corporation	1	2
NCLH	Norwegian Cruise Line Holdings	2	4
IFF	International Flavors & Fragrances	3	5
EL	Estée Lauder Companies (The)	4	9
MOS	Mosaic Company (The)	5	3
APA	APA Corporation	6	20
BBWI	Bath & Body Works, Inc.	7	10
FMC	FMC Corporation	8	1
STLD	Steel Dynamics	9	6
ON	ON Semiconductor	10	7
FCX	Freeport-McMoRan	15	8

The main empirical result of this evaluation is that 9 out of 10 firms overlap between the mean and median top 10 rankings. This proofs that there is a very high similarity between both aggregation methods since only one firm doesn't coincide. Consequently, this result shows that there is a stable risk identification across methods and at the final ranking is not driven by the arithmetic mean alone.

The very high overlap between the methods is evidence for a strong robustness in the model. The ranking is not sensitive to one extreme drawdown, volatility or crash percentile. Additionally, it shows that the multi-dimensional score produces stable results.

From the table it is observable that some rank positions change, so the ranking order is not identical for both ensemble scores. It is noticeable though, that the firm membership is largely stable, just the exact placement order differs, but the high-risk classification remains consistent.

7.5 Conclusion

Recapitulating, in this chapter the three downside risk dimensions were integrated, by using the best performing prediction model for each dimension. To propose a complete decision-making framework, the predictions were converted into percentiles and normalized, to make them comparable, by creating a common 0 to 100 risk scale. Through this, a multidimensional downside risk or model was developed.

An equal weight specification is used to achieve a balanced treatment of all downside risk measures. Specific weights can be chosen though based on necessities by investors.

The framework is able to successfully generated firm level rankings an identifying the companies with the highest risk within the S&P 500 sample. This is a very practical screening tool for investors and risk manager when deciding which companies need more attention.

As the result, Albemarle is identified as the highest risk firm under the mean ranking, which is the primary used one. Overall, the top ranked firms exhibit an elevated downside exposure across multiple dimensions. Through this it is made sure that the downside vulnerability is not based off on one specific risk measure.

Even though the arithmetic mean is used as the primary aggregation method, the median aggregation is used to check the robustness off the evaluation. The results show that there is a strong overlap between rankings since 9 of 10 firms appear in both of the top ten lists. This shows that the result is not driven by the aggregation choice but that the risk identification is stable.

Through this framework predictive models are translated into an actionable output. It can be used for investment screening, Portfolio monitoring or risk management. The emphasis is though always on ranking rather than on the exact forecasting of numbers.

8. Limitations

8.1 Cross-Sectional Design and Time Horizon

This analysis consists of a single outcome year, namely 2024. Since it is a cross-sectional analysis across firms, it doesn't have a longitudinal dimension. Because of this, the analysis has no ability to study the persistence of downside risk throughout multiple years. Consequently, the findings are conditional on the specific market environment and relationships may differ during recessions or crisis.

Future research points could be to include a panel data setting and analyse downside risk throughout multiple years. This way, dynamic downside risk predictions could be achieved.

8.2 Construction of the Crash Indicator

The crash indicator is based on the bottom 10% of annual returns. This threshold is chosen to identify the extreme negative outcomes, but the 10% cut-off ultimately remains a modelling choice. Subsequently, different thresholds could generate different classifications and firms that are close to the cutoff could switch categories. Furthermore, the crash indicator is a binary variable which contains reduced information compared to continuous outcomes.

Future research could include alternative crash definitions or multiple percentile thresholds, as well as continuous tail risk measures.

8.3 Data Quality and Measurement Error

The analysis undertaken in this thesis is based on multiple external data providers. ESG scores are obtained from a third party database and financial information is collected from publicly available sources, like Yahoo Finance. The latter though, may contain in accuracies and the ESG ratings can differ across providers. For this reason, there may be potential reporting errors in firm level variables and ESG scores. The data cleaning done before the analysis reduces but cannot eliminate measurement error.

It is important to note, that the results depend on the quality of the underlying data, so in case it contains inaccuracies, this can alter the prediction models.

Future research can consist of comparing the ESG scores across multiple providers and use alternative financial databases to check the robustness across different providers.

8.4 Model Scope and Computational Constraints

The analysis is restricted to two machine learning models, LASSO and Random Forest. Other models are excluded in this thesis. Consequently, alternative algorithms may yield different results due to their different computational methods and separate strengths and weaknesses.

Future research could be done including additional machine learning models like XGBoost, LightGBM, neural networks and a larger hyperparameter searches.

8.5 Mechanistic Interpretability Limitations

In Chapter 5, that mechanistic interpretability analysis is based on the ablation framework, where variable importance is explained through future removal. The framework is able to identify a predictive dependence, but it does not establish causality. This has a consequence, that future interactions can remain hidden and the interpretation is based on the observed performance. Mainly, ablation is able to reveal which variables matter, but it does not fully explain how the models internally combined information.

Future research could include simulation-based experiments and synthetic data environments, as well as SHAP interaction values or deeper explainable AI techniques.

8.6 Generalizability

the sample is restricted to the S&P 500 firms, which focuses only on large US corporations. These results may not generalise two small cap firms or two emerging markets. Additionally, the geographic setting is specific to the United States, so firm characteristics and ESG risk relationships can differ internationally.

Future research could consist of international samples, analyses of the European market, emerging market firms or indexes of smaller firms.

8.7 Conclusion

The limitations priorly mentioned do not invalidate the findings of this thesis, they help the reader to interpret the findings within their empirical setting of the analysis. The identified limitations serve as ideas for future research.

The framework used in the analysis remains useful despite its constraints and the findings contribute to the downside risk prediction literature.

9. Conclusion

The objective of this thesis is to understand whether firm level downside risk can be predicted using firm characteristics and to identify the main factors that are important in explaining such risk. Traditional measures often focus on average volatility, while investors are mostly concerned with adverse outcomes like severe losses, downside fluctuations or crash events.

For this reason, the analysis focuses on the S&P 500 firms and examines three distinct dimensions of downside risk: maximum drawdown, downside volatility and the crash indicator. No single risk measure can independently capture all aspects of downside exposure. This is why a multi-dimensional perspective is adopted in this thesis.

A machine learning evaluation is used through the application of LASSO and Random Forest, including this way in the analysis a linear and non-linear model.

The main findings of this thesis consist in the predictability of downside risk, where a meaningful predictive performance is achieved. It is proved that machine learning can capture firm level risk patterns.

Random Forest is found to be the strongest predictable model for maximum drawdown and the crash indicator, while LASSO wins for the predictability of downside volatility.

Volatility is consistently the most influential predictor and while ESG variables moderately contribute, the financial variables remain the dominant predictors. This can be economically interpreted such that downside risk is linked to firm characteristics. The market-based variables are the most important ones, followed by observable sector differences and ESG contributions.

To conclude the analysis, a multi-dimensional risk framework is developed through percentile normalisation and ensemble aggregation. An equal weight risk score is created, in order to rank firms based on their risk. This analysis is robustness checked through a median ranking comparison, of which the result shows a 9 out of 10 firm overlap, making it a practical decision-making tool.

Concluding, downside risk is inherently a multi-dimensional measure, so no single analysis is sufficient. Through the combination of a machine learning approach joint with an interpretability framework, the evaluation achieves substantial results and high predictability. Moreover, this framework allows the transformation from predictions into actionable insights.

This thesis contributes to the downside risk and machine learning literature, demonstrating that downside risks can be meaningfully predicted and interpreted through this multi-dimensional framework. It provides investors with a practical tool that can identify the most vulnerable firms in the market.