University of Zurich[UZH]

MASTER THESIS – Communication Systems Group, Prof. Dr. Burkhard Stiller

Quantitative-Automated Risk Consulting in the Cybersecurity Domain

Karin Brunner
Zürich, Switzerland
Student ID: 15-473-630

Supervisor: Dr. Muriel Franco, Fabian Künzler
Date of Submission: January 14, 2025

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

Master Thesis
Communication Systems Group (CSG)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
URL: http://www.csg.uzh.ch/

Zusammenfassung

Um den wachsenden Risiken in der Cybersicherheit die Stirn zu bieten, können Unternehmen Cyber Risk Frameworks und spezialisierte Modelle zur Quantifizierung von Cyberrisiken nutzen. Diese liefern bereits plausible Vorhersagen zu unternehmensspezifischen Cybersicherheitskosten, ihre Genauigkeit und Robustheit könnte jedoch durch die Einbindung zusätzlicher Daten weiter verbessert werden (z. B. tatsächliche Cybersicherheitskosten verschiedener Unternehmen oder makroökonomische Daten zu Kostensteigerungen). Cyber News sind ein Beispiel für zusätzliche Daten, die den Prozess der Vorhersage und Quantifizierung von Cyberrisiken verbessern können, da sie Einblicke in verschiedene Kategorien von Cyberereignissen geben, die ein Unternehmen betreffen und beispielsweise dessen Marktleistung beeinflussen könnten.

Da Aktienkurse in effizienten Märkten schnell auf neue Informationen reagieren, könnte der durch Cyberereignisse verursachte Gesamtverlust potenziell durch die Analyse von Veränderungen in Aktienkursen und damit in der Marktkapitalisierung quantifiziert werden. Basierend auf dieser Idee konzentriert sich diese Arbeit darauf, zu analysieren, wie Aktienkurse auf Cyber News reagieren. Hierfür werden zwei unterschiedliche Szenarien evaluiert (lineare Regression und Random Forest), um Aktienkurse auf Basis von Cyber News Daten vorherzusagen. Um die Beziehung im Detail zu untersuchen, wurde der Einfluss verschiedener Cyberereignisse auf die Aktienkurse innerhalb eines Zeitraums von drei Tagen nach dem Ereignis analysiert.

Die Ergebnisse zeigen, dass die Random Forest Regression ein leicht besseres Ergebnis erzielen konnte als die lineare Regression bei der Vorhersage von Marktreaktionen und Aktienkursänderungen, basierend auf der Menge und Sentiment-Analyse von Cyber News. Die Gesamtergebnisse blieben jedoch unbedeutend, da die Random Forest Regression einen maximalen R^2-Wert von 0.17 erreichte, während die lineare Regression 0.008 erzielte. Dies zeigt, dass beide Regressionen nur einen kleinen Teil der Veränderungen in Aktienkursen erklären können. Zudem waren die Ergebnisse über verschiedene Zeiträume hinweg inkonsistent. Die Analysen wurden mit einem begrenzten Datensatz durchgeführt, jedoch kann der Ansatz auch auf reale Datensätze angewendet werden.

Keywords: Cyber Security, Cyber News, Cyber Risk, Cyber Value at Risk, Stock Market

Abstract

To address the growing risks in cybersecurity, companies can leverage cyber risk frameworks and specialized models for quantifying cyber risks. While it already provides plausible predictions on firm-specific cybersecurity costs, its accuracy and robustness could be further improved by incorporating additional data (e.g. on actual cybersecurity costs incurred by various companies, or macroeconomic data on cost increases). Cyber news is one example of additional data that can improve the cyber risk prediction and quantification process since it can help to provide insights into various categories of cyber events related to a company that may impact, for example, companies' performance in the market.

Since stock prices in efficient markets react quickly to new information, the total loss caused by cyber events could potentially be quantified by examining changes in stock prices and, consequently, market capitalization. Based on this motivation, this work focuses on analyzing how stock prices react to cyber news. For that, two different scenarios (i.e., using linear regression and random forest) predict stock prices using cyber news data. To explore the relationship in detail, the impact of different cyber events on stock prices was examined to a timeframe of three days after the event.

As a result, the random forest regression slightly outperformed the linear regression in predicting market reactions and stock price changes due to the amount and sentiment analysis of cyber news. However, the overall results were insignificant as the random forest regression reached a maximum R^2 Score of 0.17, and the linear regression 0.008. This shows that both regressions are only able to explain a small fraction of the changes in stock prices. In addition, the results were inconsistent across different time periods. The analyses were conducted using a limited dataset, but the approach can also be applied to real-world datasets.

Keywords: Cyber Security, Cyber News, Cyber Risk, Cyber Value at Risk, Stock Market

Acknowledgments

I would like to sincerely thank my supervisors Dr. Muriel Franco and Fabian Künzler for their commitment to this topic and their efforts in acquiring the data, which made this thesis possible. I am also grateful for their willingness to share their extensive expertise and experience with me. To conclude, I greatly appreciate that they took the time to discuss and continuously refine the approach with me, ensuring that the work remained educational and interesting.

Contents

Chapter 1

Introduction

The digital transformation of recent years has not only provided companies with notable advantages but also introduced risks. Saeed et al. conclude that adopting digital solutions increases cybersecurity vulnerabilities, particularly to cyberattacks and security breaches [42]. Similarly, the Global Cybersecurity Outlook 2024 by the World Economic Forum emphasizes that the interconnectivity of the digital world accumulates negative effects, impacting everyone [60].

The importance of cybersecurity can also be substantiated with statistics. A relevant organization related to cybersecurity is the Internet Crime Complaint Center (IC3) of the Federal Bureau of Investigation (FBI). IC3 collects and analyzes data on cybercrime, raises public awareness of cyber threats, and aims to protect and secure cyberspace [22]. According to their published Internet Crime Report 2023, both the number of complaints and the total financial losses in the US have increased in recent years. As shown in Figure 1.1, IC3 received 880,418 complaints in 2023, with reported damages amounting to $12.5 billion USD [12]. Furthermore, a detailed analysis of the development of financial losses due to cybercrime was conducted by Sharif and Mohammed. They analyzed various datasets to produce statistics on the number of cyberattack complaints, losses, and victims. The clear conclusion of the paper is that the number of cyberattacks and the associated losses are increasing daily, and this trend is global [52].

A particularly vulnerable group in this context are Small and middle-sized enterprises (SMEs). According to the Cybersecurity Ventures' Cybercrime Report 2023, over 50% of all cyberattacks target SMEs [6]. Factors contributing to the higher vulnerability of SMEs include limited financial resources, lack of expertise due to a shortage of cybersecurity professionals, and underestimation of cybersecurity threats [43]. Furthermore, it is challenging for SMEs to access relevant information needed to handle or protect themselves against cyberattacks. One possible reason for this lack of data availability includes the absence of regulatory requirements for SMEs to report cyberattacks [17].

However, it is crucial for companies of all sizes to be able to quantify cyber risks. Cyber risk quantification refers to the process of measuring the cyber risks to which an organization is exposed [26]. Only through accurate quantification can businesses strategically plan cybersecurity investments and make informed decisions about how much cyber risk

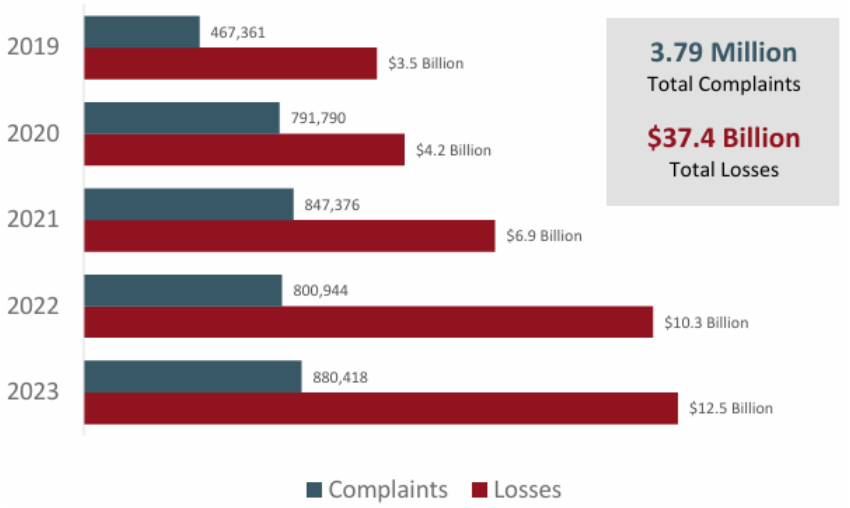


Figure 1.1: Complaints and Losses over the Last Five Years in the US [12]

a company can tolerate, how much it should mitigate, or how much it wants to transfer to an insurance provider [38].

According to the National Institute of Standards and Technology (NIST), risk is generally defined as a function of the following two components: the potential negative impact if an event occurs and the likelihood of its occurrence [34]. Therefore, if companies aim to effectively address the growing threat of cyber risks, they must develop understanding of how different cyber risks types like attacks or data breaches could negatively impact their business. At the same time, gaining insights into the likelihood of these threats is equally important. By combining the two dimensions impact and likelihood, companies can streamline their efforts and allocate resources more strategically and focus on the most critical risks.

1.1 Description of the Work

To effectively quantify their cyber risks, companies can make use of cyber risk frameworks. One example of such a framework is the Real Cyber Value at Risk (RCVaR) model, which was developed at the University of Zurich by Künzler and by Franco et al. At its core, it is a statistical analysis of cyber risk reports. The model's output estimates the potential cybersecurity costs (or losses) for a specific company that, with a given probability, will not be exceeded over a defined time horizon [29, 14]. Although the model is fully developed and has been published in academic journals, there might still be room for further improvement. Two areas that could offer potential enhancements are: (i) data availability and (ii)) the utilization of cyber risk reports.

If it were possible to access more, and especially more current, data on ongoing cyber events - rather than relying on cybersecurity reports that are several months old - the accuracy of the RCVaR model's loss predictions could be further improved. To incorporate such event data into the RCVaR model, the underlying dynamics of existing financial

theories and models such as the Efficient Market Theory (EMH) [11] could be applied. Demonstrating that cyber news data influences stock prices would allow the quantification of losses caused by cyberattacks, based on the decline in stock prices.

Based on this scenario and opportunities, this thesis has two main goals. First, it aims to analyze how cybersecurity news affects stock prices, with the analysis conducted over different time horizons and across various categories. For example, examine the impact of a cybersecurity event of day x on stock prices of days $x+1$, $x+2$, ..., $x+n$. Additionally, investigate whether the impact varies by category whether, for instance, a data breach causes a quicker and/ or more significant change in stock price compared to a cyberattack. The second goal is to use the insights from this analysis to enhance the RCVaR model. This includes gaining a better understanding of how cyber risk losses are distributed based on firm characteristics, such as company size, industry, or country, and proposing a machine learning approach to improve the predictive accuracy of RCVaR.

1.2 Thesis Outline

At the beginning of this thesis, Chapter 1 introduces the topic of cyber risk, provides an overview of its importance and lays the foundation for the research. Chapter 2 follows, offering an introduction to the fundamental concepts, which are essential for understanding and developing the topic. This chapter is divided into three sections: the first explores the RCVaR model in detail, outlining its methodology and relevance for quantifying company-specific cyber risks. The second section examines the underlying dynamics of stock prices. It provides important definitions and focuses on the factors that lead to stock price changes. The third section delves into the characteristics of stock exchanges and sheds light on the calculation of stock returns.

Based on this conceptual groundwork, Chapter 3 reviews related work in the field, and provides a detailed analysis of cyber risk frameworks and an overview of current cyber risk tools. Each of these sections concludes with a summary table that consolidates the main takeaways.

Subsequently, Chapter 4 turns to the practical implementation and describes two distinct scenarios designed to address the limited availability of cyber news data. The goal of both scenarios is to find potential connections between cyber news data and stock prices. The implementation begins with data collection, mapping, and preparation. In the first scenario, real cyber news data and actual stock prices of a specific company are used. The second scenario employs the same cyber news data but mutates it for a specific use case. Both scenarios explore potential linear and non-linear relationships between the cyber news data and stock prices.

Chapter 5 evaluates the results of these implementations, determines whether a relationship was detected in the two scenarios and compares their results. This chapter also reflects the opportunities and limitations of using cyber news data, both in general and within the specific context of this thesis.

Finally, Chapter 6 provides a summary of the findings and offers an outlook for future research. It highlights potential areas for further exploration, which includes the quantification of cyber risks, the use of cyber news data, and their influence on stock prices.

Chapter 2

Background

This Chapter 2 provides the foundational knowledge which is important for the understanding of the subsequent chapters. It begins with a detailed explanation of the existing RCVaR framework, followed by a comprehensive analysis of the underlying dynamics of stock prices, as well as an introduction to stock exchanges and a method for calculating stock returns.

2.1 Real Cyber Value at Risk (RCVaR)

The concept of RCVaR, developed by Künzler and by Franco et al., aims to accurately estimate firm-specific costs resulting from cyber risks. This is achieved by utilizing data from publicly available cyber risk reports as input for a Machine Learning (ML) model designed to predict these costs.

The primary data source is Accenture's reports, which offer comprehensive insights into various cyber incidents rather than focusing solely on data breaches. Moreover, these reports provide detailed information on the direct, indirect, and opportunity costs associated with cyberattacks and are structured on an annual basis, facilitating their application. All reports used for RCVaR are based on interviews conducted by the report publishers with employees, primarily those in key roles. Using various computer vision methods and Python scripts, information is extracted and analyzed from the cyber risk reports. The results reveal a clear tail-heavy distribution of costs, which is incorporated into the RCVaR calculation, accurately reflecting the real dynamics of these costs.

The complete calculation is then performed according to Equation 2.1, which is composed of the company valuation, a size scaler, a time scaler, and a factor scaler.

$$\text{company\_cost}_{\text{year}} = \frac{\text{valuation}_{\text{ReportYear+t}}}{\text{discount}^{t}_{\text{valuation}}} \times \text{cv\_ratio} \times \text{discount}^{t-\text{ReportYear}}_{\text{cost}} \times \prod_{i=1}^{11} (1+\text{param\_ratio}_i) \tag{2.1}$$

The first component, the company valuation, consists of the equity value of a company in the year from which the data in the cyber report is derived. In the next step, this company valuation is multiplied by the size scaler, resulting in potential costs as an initial intermediate result. The size scaler represents the average costs, calculated by dividing the average market costs by the average market capitalization. These potential costs are then multiplied by the time scaler to scale the costs to a desired year. The time scaler consists of two discount factors: the cost discount factor and the valuation discount factor. The cost discount factor is intended to replicate cumulative cost development, while the valuation discount factor accounts for valuation adjustments due to inflation. Through extensive research, the authors defined value of 9.6% for the cost discount factor as appropriate, and a value of 1.8% was determined for the valuation discount factor based on the analysis of historical inflation rates. In the final step, named factor scaler, the cost calculation is tailored to a specific company using 11 different factors. These factors include general firm-specific information, such as the industry and the country in which the company operates. Additionally, there are factors with a direct connection to cybersecurity, such as cyber insurance or security measures.

As model output, a company receives cyber risk costs in USD, representing the amount that is unlikely to be exceeded within a specified timeframe and probability. Figure Figure 2.1 provides a visualization of the RCVaR. The green area shows the distribution of expected losses with a cumulative probability of 95%. The purple section represents the losses with a low probability of 5% that exceed the reported RCVaR of USD 23,448.94.

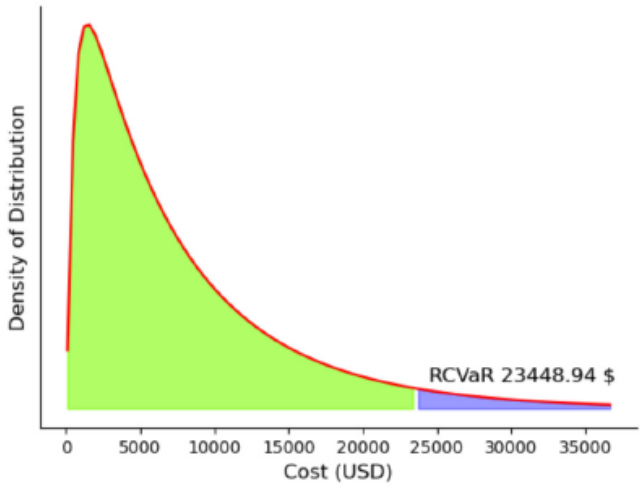


Figure 2.1: Visualization of RCVaR With 95% Confidence [14, 29]

In addition to the theoretical development of the RCVaR approach, it was also implemented in practice. The front end was developed using React and TypeScript, while the back end relies on Flask to deliver the RESTful API.

In the next step, the model was tested to draw conclusions about the accuracy of its cost and risk components. For the evaluation of cost predictions, cybersecurity reports from three providers were used, which were not part of the model's training data. These reports' information can therefore be considered as out-of-sample data. The analysis revealed that the costs predicted by RCVaR closely aligned with the actual costs from the

first report, with a difference of only 9%. However, for the other two reports, relatively larger deviations of 44% and 53% were observed. It is worth noting that these two reports provide costs per incident, whereas RCVaR estimates annualized direct and indirect costs. Therefore, the authors conclude that RCVaR has the potential to approximate costs from various reports but emphasize that further testing is necessary, particularly to validate the accuracy of the discount and cost-increasing factors.

Regarding risk assessment, the analysis demonstrated that the cost distribution predicted by RCVaR aligns with distributions reported in the literature, indicating that RCVaR can realistically represent real-world costs.

Although RCVaR provides relatively accurate and realistic predictions for cost and risk components, the results rely on certain assumptions and are subject to limitations. For example, it is assumed that the cybersecurity reports reflect the actual costs incurred and are largely unaffected by selection bias. Despite the careful preparation of these reports by their publishers, such uncertainties cannot be entirely ruled out. Moreover, the anonymization in the reports makes it difficult to associate company-specific attributes, such as valuations, with the corresponding losses. In addition, RCVaR is based on simplified economic principles, for instance, inflation is used as a valuation discount factor. With more data, it would be possible to assess whether inflation accurately represents the valuation discount over the long term. If this is not the case, suitable alternatives could be explored. A further advantage of incorporating additional data, as noted by the authors, is that it could improve the overall robustness of the results [14, 29].

2.2 Underlying Dynamics of Stock Prices

First, several important terms are defined according to the definitions provided by Lee and Lee. Stocks represent shares of ownership in a company and their market price reflects the current amount of money at which they are bought or sold in a marketplace. By multiplying the price of a stock by the number of outstanding shares, the total value of a company, known as its market capitalization, can be determined [31].

To understand what influences stock prices, the concept of efficient markets is essential. The theoretical foundations of the so-called Efficient Market Hypothesis (EMH) were examined and formally established in a notable paper by Fama. The paper defines a market as informationally efficient if stock prices at any point in time reflect all available information. Furthermore, it establishes three distinct levels of market efficiency. In the first level, the weak form, prices incorporate only historical information. The semi-strong form extends this by reflecting not only historical but also newly released public information. Finally, in the strong form of market efficiency, prices include all historical, public, and private information. Factors that generally promote market efficiency include the absence of transaction costs, free access to all information, and consensus among market participants regarding the interpretation of this information [11].

Another paper by Fama offers an intuitive explanation of why information affecting a company's outlook is always priced into stock values. When new information becomes

available, stock prices adjust immediately. This happens because market participants strive to use new information to gain a financial advantage. This advantage is exploited only until the effect is neutralized [10].

If, for example, a company reports lower-than-expected annual earnings, its stock price will immediately drop. This is because investors had anticipated higher earnings, and that public expectation was already factored into the stock price. When the company fails to deliver the expected profits, the stock price is no longer justified and will adjust downward until it has reached the appropriate level according to the investor's new assessment.

In case cybersecurity news is also considered relevant to the company's outlook from an investor's perspective, it will influence stock prices and, consequently, the market capitalization. Therefore, the total loss due to a cyber event for a company can be estimated by analyzing the difference in market capitalization before and after a cyber event. This approach inherently includes future expectations in the total losses, as the public perception of a potential negative impact on profits would lead to a downward adjustment of the stock price to reflect the new reality. Using this method, the financial losses caused by cyber events could be analyzed across a wide range of companies, varying in factors like size or industry. This would not only provide deeper insights into the distribution of losses but also help refine RCVaR and further improve its robustness.

2.3 Stock Exchanges and Returns

Stocks are bought and sold on stock exchanges, which serve as intermediaries between buyers and sellers. There are numerous stock markets worldwide, each operating in different time zones which results in distinct trading hours. It is important to note that stock exchanges are not operational on weekends, public holidays, or outside their official trading hours. This means that normally during these periods, no trading activity is possible. If a buy or sell order is placed at a time when the exchange is closed, it will be queued and executed once the exchange reopens [20].

To calculate the profit or loss from investing in a stock, there are various ways to compute the return. One such method is the Holding Period Return (HPR). According to the definition by Mondello, this measure can be applied over any time period and provides insight into the percentage change of the investment relative to its initial value. The HPR consists of the capital return and the dividend return, although, depending on the context, the formula for either the capital return or the dividend return alone can be used. HPR is a straightforward method for measuring investment performance, however, it does not account for potential cash flows during the holding period. If cash flows were present, the Money Weighted Rate of Return could be a more appropriate metric. The formula for HPR is presented in Equation 2.2 [33].

$$\text{Holding Period Return (HPR)} = \frac{(P_t - P_0) + Div_t}{P_0} = \frac{P_t - P_0}{P_0} + \frac{Div_t}{P_0} \quad [33] \qquad (2.2)$$

where:
P_0 = Price of the asset at the beginning of period t
P_t = Price of the asset at the end of period t
Div_t = Dividend at the end of period t

If, for example, the 1-day capital return of a stock is to be calculated, subtract yesterday's end-of-day price from today's end-of-day price and divide the result by yesterday's end-of-day price.

For the purposes of this thesis, the concept of Holding Period Return (HPR) is appropriate, as there are no cash flows involved. Its application offers two key advantages. First, the HPR is expressed as a percentage, which makes it independent of the absolute magnitude of a stock price and its currency. This enables a straightforward comparison across different companies. Additionally, the HPR is a simple and flexible method, and therefore, allows the quick implementation of different time horizons. This simplifies the analysis of the impact of a cyber event on stock prices over different periods.

Chapter 3

Related Work

This chapter provides an overview of potential solutions for cyber risk quantification. Section 3.1 examines both established and emerging academic frameworks, while Section 3.2 focuses on practical solutions from the cyber risk industry. The two perspectives offer together a comprehensive view of the current state of cyber risk quantification in academia and practice.

3.1 Cyber Risk Frameworks

This section provides a brief overview of academic cyber risk frameworks, with a more detailed analysis of their strengths and weaknesses. The terms "framework" and "methodology" are used interchangeably, following the definitions provided in [13]. Table 3.1 summarizes the frameworks discussed, highlighting their respective strengths and weaknesses. The column named "Approach" offers the reader an initial insight into each paper's methodology. Frameworks labeled as "Established" are those that originated at least 10 years ago and are frequently referenced as "well-known" or similar in contemporary academic literature.

An example of a widely referenced framework is the NIST Cybersecurity Framework (CSF), first published in 2014 by the National Institute of Standards and Technology (NIST) [35]. The subsequent version, NIST CSF 1.1, supports companies in expressing, managing and understanding cybersecurity risks and consists of three parts "Framework Core", "Framework Tiers" and "Framework Profile". The framework provides information on industry standards, guidelines and practices to help companies assess their current cyber risk management practices and identify opportunities for improvement in cybersecurity [36]. As NIST is a recognized organization, the framework gains a high level of acceptance and credibility. Other advantages are its flexibility and conciseness. However, it is rather difficult for smaller companies to implement [30]. In 2024, the NIST CSF 2.0, an updated version of the framework, was released [37]. This more recent version adopted the basic concepts of NIST CSF 1.1 but is more appropriate for all sizes & types of organisations [30]. Nonetheless, both versions are flexible and practical but share the

disadvantage of lacking financial risk quantification, making risk interpretation challenging. Additionally, it is a qualitative risk classification, which involves a certain degree of subjectivity.

In contrast, The Factor Analysis of Information Risk (FAIR), is a purely quantitative framework [9]. It uses a risk taxonomy, or ontology, to divide the risk into factors and finally calculates the risk as a combination of the probability of "loss event frequency" and "loss magnitude" [16]. Moreover, it can be used in combination with existing frameworks [39]. The risk quantification approach is regarded as comprehensive and the risk estimation process is complete. In addition, it offers tools for risk measurement and quantification [58]. Although the comprehensiveness of the approach is seen as an advantage, the number of different factors could overwhelm a non-expert and discourage them from using the framework. In addition, a company must have historical and current data available (e.g. on cyber incidents) to use the framework.

Another framework, whose origins date back more than 10 years is the ISO/IEC 27005 standard [24]. The first version, ISO/IEC 27005:2008, was introduced in 2008 [23]. It's current version, ISO/IEC 27005:2022, guides companies in complying with ISO/IEC 27001 to deal with information security risks and in performing risk assessments [24]. It is considered an overall complete approach as it covers aspects of risk identification, estimation and evaluation process. It is even regarded as "industry best practice". However, there are missing parts in the risk identification process and prior knowledge is needed for certain steps of the quantitative risk estimation process [58].

In addition to the established frameworks, there are numerous innovative approaches. One of these is Cyber Value at Risk (CVaR), a quantitative approach that uses the Value at Risk (VaR) concept. VaR is a proven method in the financial sector for quantifying worst-case losses over a specific period of time and within a certain confidence level [38]. Compared to VaR, CVaR focuses on the quantification of losses due to cyberattacks and indirectly considers the effectiveness of security measures [14]. CVaR was first proposed by The World Economic Forum in 2015 [59], however, they did not provide a detailed calculation methodology. To close this gap and make the CVaR concept applicable in practice, Erola et al. [8] developed a concrete methodology. It utilizes Monte Carlo simulations to estimate potential financial losses from cyber incidents. To do so, data from Cyence was utilized to assess the effectiveness of controls, while data from Advisen provided information on threats and their probabilities. Additionally, company-specific infrastructure, the effectiveness of implemented controls and the connection between particular threats and assets are taken into account. By calculating losses through simulations rather than aggregating datasets, the company's assets are evaluated comprehensively, free from dataset limitations [8].

Since the proposed methodology is based on the CVaR concept and represents a concrete application of it, it inherits its advantages and disadvantages. The advantages include that the estimation's output is a monetary value, making it easy to understand and interpret. Additionally, it can be compared with the VaR from other domains. However, CVaR fails to properly account for rare, large-scale events [14]. Another drawback of CVaR is that it is a backwards-looking measure, as it relies on historical figures and assumes that past assumptions will hold true in the future [40]. What also adds to this is that

the use of simulations comes with some general drawbacks. They introduce a certain level of complexity and require large amounts of data [27]. When experts are involved in addressing these complexities, their personal experiences can introduce biases into the process [14]. To sum up, the proposed methodology provides an easily interpretable and company-specific output. However, it relies on non-publicly available data and inherits various drawbacks from both the VaR and simulation approaches. Additionally, there is a dilemma concerning company-specific data. On the one hand, it is beneficial to use internal company data to incorporate information on infrastructure, controls, and threats. On the other hand, this requires the company to have the knowledge and infrastructure to collect and potentially store the data. Furthermore, having access to historical data would be useful to ensure a large dataset, which makes the process even more complex and resource-intensive.

A novel approach that aims to leverage the advantages of CVaR while addressing its drawbacks is RCVaR. It is designed to help companies estimate the financial impact of cyberattacks. Like CVaR, RCVaR is expressed in financial terms, therefore, easy to understand, interpret and compare. What sets RCVaR apart is its use of publicly available real-world data, enabling companies without an extensive historical database to effectively assess their cybersecurity risk. Moreover, no statistical or cybersecurity expertise is required to apply the method and it is suitable for companies of any size. It also takes into account long-term costs, such as reputational damage, ensuring a more comprehensive risk assessment. By taking full advantage of these benefits, it enables decision-makers to find the optimum between costs and benefits of security investments. Another noteworthy advantage is that this approach can provide estimates for both future and past periods. However, despite the numerous advantages, RCVaR faces certain drawbacks, including the possibility of data pollution and cross-correlation, which can affect the accuracy of the estimations. Additionally, the use of publicly available reports can introduce selection bias, and the anonymization within these reports makes it difficult to assign a valuation to each company. Therefore, the RCVAR cost estimation is sensitive to the valuation input. Lastly, it is important to note that the discount factors play a crucial role in the estimation of the market capitalization or costs are scaled over the years [14, 29]. A detailed description and discussion of the approach can be found in Section 2.1 of this Master Thesis.

Another innovative approach, however, without any connection to VaR, is Yet another cybersecurity risk assessment framework (Yacraf). It is a model-based approach that combines components from threat modeling and quantitative risk assessment and builds on the well-known frameworks PASTA and FAIR. From the former, it utilizes the concept of Data flow diagrams (DFDs), while from the latter, it further details and expands the vulnerabilities assessment. This assessment includes qualitative components. The final results include financial losses per attack event, therefore enabling a transparent cost vs. benefit analysis [7]. Knowing the loss per attack event is valuable and certainly helpful information for companies, supporting a transparent and informed decision-making process. However, the qualitative aspects of the vulnerability assessment mean that either prior knowledge is required or experts must be consulted. Additionally, the qualitative elements can introduce subjectivity into the process.

The frameworks presented so far have primarily been developed with a focus on larger

companies or applicable to businesses of all sizes. Sukumar et al. [55], however, chose a different focus and developed a framework specifically for SMEs. The developed multilevel decision-making framework combines two existing techniques Step-wise Weight Assessment Ratio Analysis (SWARA) and Best-Worst Method (BWM) in order to rank cyber risks. It offers several advantages, particularly its suitability for SMEs, and assists companies in prioritizing their cybersecurity investments. Furthermore, drawing from two existing methodologies brings together their strengths to create an approach to managing cybersecurity risks. On the downside, the framework is dependent on the judgment of experts. Additionally, rare but impactful events can strongly influence the overall assessment [55]. This reliance on specialists introduces the potential for bias or incomplete assessments if vulnerabilities are overlooked or their importance is misjudged. Consequently, these factors can negatively affect the reliability of the risk prioritization.

Another interesting approach suitable for SMEs is offered by Gandal et al. [18]. The approach aims to quantify and reduce cyber risks by analyzing the relationship between vulnerabilities, security measures and incidents. The insights gained should help companies determine which cybersecurity investments are most worthwhile. For this approach, a remarkable dataset was created using solely publicly available data. Thanks to this dataset, real-world information on vulnerabilities, email attacks, incidents, and precautions at company level was incorporated into the analysis. However, it should be noted that the insights gained are based on a small dataset and would need to be validated with a larger sample size. Additionally, there may be issues with endogeneity, as panel data were not used [18]. Furthermore, it is not fully explained how a company can calculate the individual costs of having at least one security incident are calculated. While the formula is defined as "probability of incident * expected loss", the explanation for the "expected loss" component is missing. Although the paper refers to a mean loss per cyber incident of 36.8 million Euros [4], this figure, unlike the probability, is not based on firm-specific data of SMEs.

Jiang et al. [25] describe an approach that also takes advantage of the vast amount of publicly available data. ML techniques are used to predict the company-specific probability of cyberattacks and thus assess the cyber risk. The data used comes from CRSP, Compustat, the Identity Theft Resource Center (ITRC), and from companies' 10-K filings. By incorporating the "Item 1A: Risk Factors" of the 10-K filings, a forward-looking self-assessment by the companies, is integrated into the model. However, the method does not provide any indication of the potential scale of losses that could occur [25]. The main advantages of this method include the fact that a company doesn't need to provide internal data to use it. Additionally, it provides insights into future cyberattacks, offering valuable and easy to understand information. However, a significant drawback is its reliance on 10-K reports. As 10-K reports are only required for U.S. public companies [51], the method is limited to this set of companies.

For organizations seeking information on the scale of loss or those without a 10-K report, the Quantifying Cyber Risks for Strategic Decisions (QBER) approach could be useful. It quantifies cyber risks by combining technical, economic, and legal perspectives. This enables companies, in addition to assessing their financial risks, to gain more insight into the current state of compliance with regulations (legal risks) and internally implemented precautions (technical risks). Both publicly available data, such as industry reports,

and internal data, such as information on implemented controls, are incorporated into the calculations. This data is then combined with analysis from security specialists, allowing the probability and risks of cyberattacks to be quantified. QBER is a holistic and comprehensive approach that supports companies in their cybersecurity strategy. This allows decision-makers to make informed decisions and choose solutions with the best cost vs. benefit ratio. Additionally, the approach is flexible and can be applied to various scenarios, regulations, and precautions [15]. To apply QBER, companies need to provide internal data as input. This could pose a challenge, particularly for smaller companies or those with limited cybersecurity knowledge. Additionally, the opinions and expertise of security experts are integrated into the process, which introduces a certain level of subjectivity and could lead to biases or misjudgments.

When looking at Table 3.1, it quickly becomes clear that no method offers only advantages. It is, therefore, impossible to identify a single approach as the best. However, Table 3.1 also shows that there are many different approaches, each offering its specific benefits. For example, frameworks like [14] and [25] are forward-looking and able to predict future risks or costs. The frameworks [14], [7], [55], and [15] allow for transparent cost-benefit analyses or prioritization for cybersecurity investments, helping to make the best possible decisions. Another positive feature is when the risk is expressed in financial terms, as this is easier to understand and interpret. This is achievable with the frameworks discussed in [14], [7], and [25]. Moreover, a framework should be broadly applicable, such as for SMEs, as seen with the frameworks [37], [14], [55], and [18]. In addition, a framework should be as holistic as possible, taking into account a wide range of factors, which framework [24] manages excellently. The QBER framework [15] is another example, even incorporating the legal perspective. A common drawback shared by many frameworks is the need for cybersecurity knowledge to apply or their input is needed for certain parts of the process. For instance, frameworks like [24], [8], [7], [55], and [15] require specific expertise or experience in cybersecurity or statistics. The use of internal data presents a double-edged sword. On the one hand, the company-specific situation regarding infrastructure and controls should be considered as much as possible to provide an accurate prediction. However, from a company's perspective, it is desirable to minimize the use of internal data, as it may not be (historically) available, or experts may be needed to prepare the data. Only the frameworks [8] and [14] offer a solution to this problem by indirectly including these factors in the process.

In conclusion, it can be stated that the framework defined by RCVaR [14] stands out due to its many positive qualities. It is future-oriented, applicable to SMEs, usable without specialized knowledge, and produces easily understandable outputs. Additionally, it helps in cost-benefit evaluations and does not require companies to provide internal data. However, even the RCVaR has limitations, leaving room for enhancements and further development. A discussion about these limitations can be found in Section 2.1 of this thesis.

Furthermore, no existing framework was found that utilizes news data related to cybersecurity events, nor does it combine news data with stock prices. Most examined models focus primarily on historical data or specific cyber incidents and do not consider current dynamics for risk quantification. Therefore, the identification of a relationship between news data and stock prices, as well as integrating this into RCVaR, would not only repre-

sent a unique and novel contribution but also offer valuable insights for companies, which aim to proactively manage cyber risks.

Table 3.1: Overview of Cyber Risk Frameworks

Framework	Year	Approach	Established	Strengths	Weaknesses
NIST CSF 1.1 [36]	2018	Qualitative risk classification	✓	Flexible; concise; trustworthy issuer	Challenging for SMEs; no financial risk quantification; potential subjectivity
NIST CSF 2.0 [37]	2024	Qualitative risk classification	✓	Flexible; concise; trustworthy issuer; considers all sizes & types of organizations	Seems initially more complex than NIST CSF 1.1; no financial risk quantification; potential subjectivity
FAIR [16]	2015	Quantitative risk taxonomy	✓	Complementary to existing risk frameworks; comprehensive risk quantification; complete risk estimation; provides tools for risk measurement and quantification	Knowledge of different factors required; company internal data needed
ISO/IEC 27005 [24]	2022	Qualitative information security risk assessment	✓	Complete risk assessment process; industry best practice	Missing parts in risk identification process; prior knowledge of statistics needed
Erola et al. [8]	2022	Monte Carlo simulation-based quantitative risk model	×	Result is easily interpretable; uses well known economic concept; cross-domain risk comparison; includes effectiveness of risk control measures; holistic view on company assets	Used datasets not for free; lack of data for real-world use; backwards-looking measure; rare huge events are not well represented; simulation approach: uses large amounts of data and possible bias due to involved experts
Franco et al. [14], Künzler [29]	2024	Quantitative data-driven economic risk model	×	Based on public real-world data; usable for non-IT-experts; result is easily interpretable; uses well-known economic concept; finds optimum between costs and investments; usable for SME; cross-domain risk comparison; long-term costs are considered; predictions are possible	Possibility of selection bias, data pollution, cross-correlation; anonymization in reports is challenging; huge dependency on market capitalization and discount factor
Ekstedt et al. [7]	2023	Model-based cybersecurity risk assessment framework	×	Extends well-known frameworks; transparent and structured decision support; financial loss per attack event available	Partially qualitative; knowledge required; potential source of bias
Sukumar et al. [55]	2023	Multi-criteria decision analysis (MCDA) method	×	Usable for SME; helps to prioritize cybersecurity investments; combines strengths of existing approaches	Depends on experts' opinion; rare huge events influence assessment
Gandal et al. [18]	2020	Statistical model	×	Based on public real-world data; usable for SME; includes firm specific measured risk and precautions data	Based on small sample; endogeneity problems; cost estimation not fully company specific
Jiang et al. [25]	2024	Machine learning-based risk assessment framework	×	Forecasts cyberattacks; forward-looking; no internal data needed	Focus on likelihood not scale of losses; only for listed U.S. companies
Franco et al. [15]	2024	Quantitative, multidimensional cyber risk assessment framework	×	Flexibility; includes scenario analysis; offers decision support; includes legal aspects	Internal data needed, expert knowledge required

3.2 Current Instrustry Cyber Risk Solutions

Cybersecurity threats pose a significant risk to companies of all sizes and locations. To address and manage these risks systematically, a cyber risk management process is essential. According to IBM, this process consists of the following steps: framing, assessment, response, and monitoring. In the first step, risk framing, the scope of the risk is defined. This includes identifying which systems and data may be exposed to threats and assessing their importance. The second step, risk assessment, involves identifying vulnerabilities and potential threats, and evaluating their potential impact on the company. In the third step, risk response, strategies are implemented to address the risk. For example, risks can be mitigated or transferred through insurance. The final step, risk monitoring, involves continuously monitoring systems and the threat landscape to quickly identify new vulnerabilities [21]. Steps 1 and 2 primarily involve cyber risk planning, as they focus on identifying company-specific assets, threats, vulnerabilities, and potential impacts, as well as comparing possible strategies for addressing these risks. Steps 3 and 4 then focus on implementing the measures planned in the earlier stages.

Companies looking to optimize their cybersecurity efforts, particularly in the area of cyber risk planning, have a wide range of tools available on the market. The next section presents a selection of these tools, while Table 3.2 provides an overview, including potential use cases and an own classification of the approach.

The first tool presented by Zeron provides an all-in-one dashboard for cyber risk management, as shown in Figure 3.1. It uses AI to centralize and standardize data related to cybersecurity, quantifies the cyber risk posture through a risk score and thereby provides valuable insights. Besides a cyber risk posture assessment, it additionally offers attack surface and defense automation. This ensures that potential vulnerabilities are detected immediately and the effectiveness of cybersecurity measures is continuously evaluated. Another useful feature is compliance automation, which helps ensure adherence to standards and regulations [63].

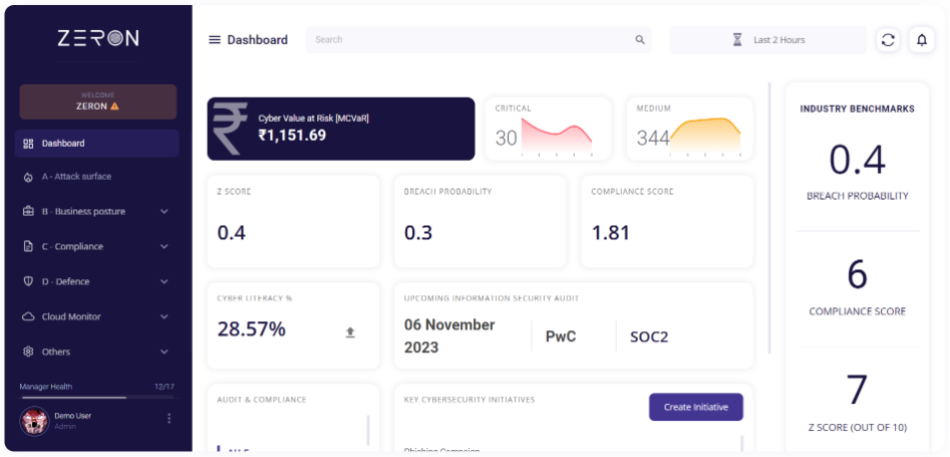


Figure 3.1: Zeron all-in-one dashboard [62]

Alternatively, Kovrr also offers a dashboard, but its approach does not rely on AI. Instead, the Monte Carlo simulation-based cyber quantification platform leverages company-

specific, cyber insurance, and threat data to perform quantitative cyber risk analysis. It calculates the potential financial loss across different scenarios while considering various business areas. For instance, in the event of a data breach, the platform accounts for losses from areas such as legal fees and revenue loss. Additionally, it tracks how company-specific cyber risks evolve over time and highlights the financial savings achieved through the implementation of recommended mitigation measures. One such recommendation could be upgrading to a particular framework, such as NIST. Furthermore, the platform includes an ROI calculator that facilitates cost-benefit analyses for potential cybersecurity investments. Notably, it also allows companies to compare their cyber risk against industry benchmarks [28].

RiskLens Enterprise provides an additional approach for carrying out scenario analyses. It is a Software as a Service (SaaS) platform with a focus on easy and quick cyber risk quantification. As the creator of the Factor Analysis of Information Risk (FAIR) standard, RiskLens calculates risk in accordance with the FAIR taxonomy. The platform also provides automated guidance for applying the FAIR standard and streamlines the collection and storage of cyber risk data. This data is used to generate standardized reports. In addition, the platform simplifies scenario modeling, making it straightforward and intuitive to explore and evaluate various risk scenarios. An API is available, enabling users to import data or export analysis results [41].

X-Analytics is a tool without explicit API integration, but it offers a range of other helpful features. It uses open-source and industry data to help companies prioritize and communicate cyber risks, as well as protect against them. In particular, it aids in understanding risk exposure by highlighting the financial impact across various areas, such as data breaches or ransomware, enabling targeted risk prioritization. Additionally, 'what-if' scenario analyses can be performed, and reports and ROI visualizations can be generated [61].

The Risk Quantifier (RQ) tool from ThreatConnect is another option for companies seeking to assess the financial impact of cyber risks. It is AI-powered and as the name suggests, its focus is on risk quantification and offers data, analytics, and recommendations to elevate a company's cybersecurity posture. The tool can be used to prioritize risks and mitigation measures, ensure compliance with regulatory standards such as SEC Materiality, and conduct risk exception and gap analyses [56].

All the tools presented so far are geared towards cyber risk management or cyber risk quantification. SOCRadar, however, takes a different approach by focusing on (cyber) threat intelligence. This term refers to detailed information about currently relevant cybersecurity threats, which are promptly analyzed to support decision-making and enable targeted actions [1]. SOCRadar provides functionalities to achieve this, such as updates on threat actors and insights from the dark web, including information on breached datasets. It also offers advanced alerts for critical vulnerabilities. Beyond Threat Intelligence, SOCRadar supports Attack Surface Management, including monitoring digital assets and SSL certificates [54]. These features are more aligned with risk response management than traditional cybersecurity planning. The insights gained from the threat landscape and the dark web provide companies with valuable input for cyber risk planning.

PREACT security optimization platform by AttackIQ offers a different approach to tackling cyber risk planning. Continuous automated testing of implemented measures helps identify a company's vulnerabilities, enabling better decision-making regarding technology, processes, and workforce. Additionally, it provides real-time visibility into the current status of all assets and control mechanisms, ensuring that up-to-date information is integrated into the planning process [3].

The tools presented in this section share some common features in their design. Typically, there is an initial step in which relevant internal or external data is collected, such as internal information on (real-time) vulnerabilities, assets, or implemented measures, and external information on the current cyber threat landscape. Depending on the tool's approach, the next step involves running various risk or loss scenarios and/or prioritizing risks to identify areas with the greatest need for action. Based on the insights gained, recommendations are made on how to improve the situation, sometimes supported by ROI calculations. Some tools also include suggestions for regulatory compliance. A further commonality is the final step, where results are presented in a visually appealing way, either in a dashboard or report. This process provides companies with comprehensive information about their current cybersecurity status, enabling them to plan and establish their strategy accordingly.

All tools feature user-friendly interfaces and appear to address the needs of companies well, offering assistance in managing cyber risks using the latest technology. However, none of the websites mentions independent tests to verify the shown scenarios, impacts, or risk prioritizations. Instead, [28], [61], [63], [56], and [3] list companies on their websites that use their tools. However, there is no information on how extensively these tools are used, how satisfied the companies are overall with the tools, or with the accuracy of the calculations. Additionally, the calculation methodologies are not disclosed in detail, presumably to protect the uniqueness of the solutions. This lack of transparency makes it difficult to fully understand the tools from the outside, giving them the appearance of a 'black box.'

Therefore, approaches and tools with a highly transparent calculation methodology backed by data-proven accuracy remain necessary. Furthermore, additional quantitative and qualitative evaluations are needed in the field to highlight the measurable benefits of cybersecurity tools for risk quantification.

Table 3.2: Overview of Current Industry Cyber Risk Solutions

Provider	**Approach**	**Use Cases**
Zeron [63]	AI-powered cyber risk quantification	• Real-time identification of vulnerabilities • Cybersecurity posture assessment • Automated compliance with standards and regulations • Continuous evaluation of internal measures' effectiveness
kovrr [28]	Monte carlo simulation powered cyber risk quantification	• Loss scenario analysis • Risk progression overview • Risk mitigation recommendations and risk industry comparison • ROI calculator
RiskLens [41]	Monte carlo simulation powered cyber risk management	• Automated guidance for implementing FAIR • Straightforward analysis of loss exposure across multiple scenarios • Simplify data management and standardized reports • Easy data import and export
X-Analytics [61]	Data-driven cyber risk management	• Understanding cyber risk exposure • Scenario analysis and risk ranking • Reporting and ROI visualization
ThreatConnect [56]	AI-powered cyber risk quantification	• Risk and measures prioritization • Compliance with regulatory standards • Risk exception and gap analysis
SOCRadar [54]	AI-powered cyber threat intelligence	• Detailed information about the cyber threat landscape • Information on threat actors and from the dark web • Advanced vulnerability detection
AttackIQ [3]	Breach and attack simulation-based security control validation	• Real-time data on control's effectiveness • Investigate the asset's current state • Decision support on technology, processes, personnel

Chapter 4

Approach

This chapter focuses on the design and implementation of an approach to explore the relationship between stock prices and cyber news data. The objective is to assess whether cyber news data can be used to predict stock prices. If a relationship exists and is sufficiently strong, it should enable accurate predictions of stock prices. To investigate this, both a linear and a non-linear machine learning (ML) model are implemented.

The implementation begins with gathering cyber news and stock price data from two different sources. The data is then analyzed to understand its scale, distributions, and to identify any unique characteristics. Following this, the data is mapped and prepared for further use. Subsequently, a linear and a non-linear ML model are selected as suitable choices for the given approach. These models are then implemented, and in the final step, a brief analysis of the results is conducted.

The work relies on data from Apple Inc. (Apple) and generates synthetic data based on market behaviors. This is due to the fact that data sharing is challenging within the cybersecurity field. The scarcity of data is also a problem for the cybersecurity economics field since it is an emerging topic, and there are few companies that handle this kind of data. Therefore, despite significant efforts during this Master Thesis, it was impossible to resolve bureaucracy and details surrounding the data usage to enable access to complete real-world datasets to develop the thesis.

To adapt to these limitations, the approach was refined to include two distinct scenarios, as illustrated in Figure 4.1. The first scenario focuses exclusively on real data, aiming to examine whether a relationship exists between the cyber news data for Apple and Apple's stock price. This scenario is detailed in Section 4.2.

In contrast, the second scenario includes synthetic data. While real share prices are still used, Apple's cyber news data is manually adjusted to create synthetic inputs. This approach investigates whether a connection exists between the modified cyber news dataset for Apple and the stock price of another company. A comprehensive discussion of this scenario can be found in Section 4.3. To provide clarity, Figure 4.1 clearly labels the steps involving modified data as "synthetic," distinguishing them from those using real data. This ensures a transparent understanding of the data sources applied in each part of the approach.

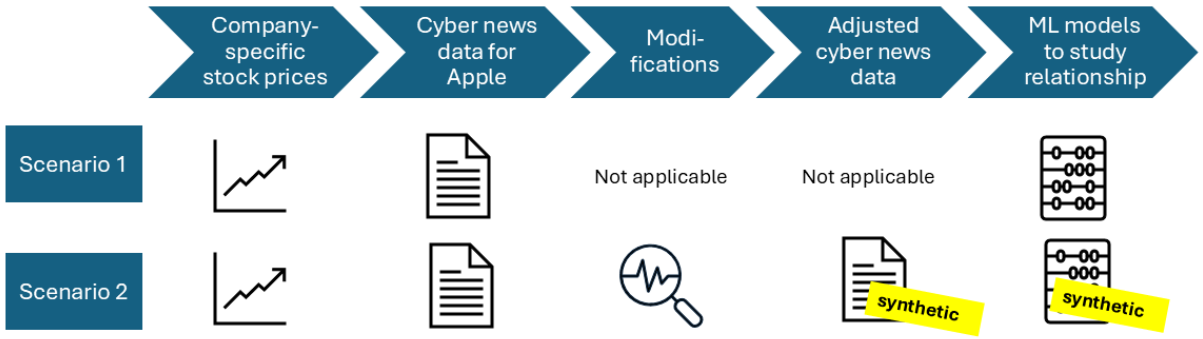


Figure 4.1: Visualization of the New Approach

It is also important to highlight that the code developed for the data pipeline and the ML models is reusable and can be applied beyond the scope of this thesis. Once access to the complete dataset becomes available, the prepared code will enable a comprehensive analysis of the full dataset, allowing the approach to be expanded as needed. The complete code can be found at the following link:

https://github.com/kb524/Cybersecurity_Risk_Consulting.git

Although the full code can be found in this GitHub repository, only the synthetic cyber news data is available. The real cyber news data cannot be shared due to an NDA.

4.1 Data Gathering and Preparation

Two different data sources are used for the implementation and validation of the proposed approach. The first source consists of cyber news data, while the second comprises publicly available stock prices.

The cyber news data is provided per company in three separate CSV files, each containing the date and one to four additional data fields. Table 4.1 provides an overview of these data fields along with a categorization.

Table 4.1: Overview of Cyber News Datafields

Name	**Type**	**Category**
Date	Datetime	Event date
Cyber Attack	Integer	Event type
Cybersecurity	Integer	Event type
Data Breach	Integer	Event type
Data Security Management	Integer	Event type
Volume of News	Integer	Events total
Perc. of Positive Sentiment	Float	News sentiment

The first data field ”Date” identifies the specific date linked to the event and sentiment data. The subsequent four data points provide information about the events ”Cyber Attack”, ”Cyber Security”, ”Data Breach” and ”Data Security Management”. These data

are generated daily at the end of the day for each company by the data provider. They are verified by an independent news source to ensure that only confirmed events are included in the dataset. These events are not classified as positive or negative but are simply associated with a specific company. The data point "Volume of News" represents the total number of event data entries per company on a given day. It is important to mention that this column is not the sum of the mentioned cyber event type columns as other news events are part of this column.

The final data point "Perc. of Positive Sentiment" provides insights into whether the sentiment of a specific date can be interpreted as positive or negative. The overall sentiment is determined by dividing the number of positive mentions by the total number of mentions for the company. A value of 0.5 is considered neutral. Values above 0.5 indicate a positive sentiment, while values below 0.5 suggest a negative sentiment. The explanations regarding the cyber news dataset were provided by one of the supervisors.

The implementation is primarily carried out using the Python programming language. For data gathering and processing, the packages pandas and os are utilized. In the initial step, the individual CSV files are loaded and converted into separate pandas DataFrames. Additionally, a column containing the company identifier ISIN is added, which will be used for later analysis and mapping. Following this, the three DataFrames per company are merged, and the resulting DataFrames from all companies are combined into a single comprehensive DataFrame.

In contrast to the cyber news data, the required financial data is not provided in CSV format. To obtain the daily historical stock prices, the LSEG Datastream login provided by the University of Zurich is used [57]. LSEG is a provider of financial markets infrastructure, data, news and indices [32]. Access is granted through the Excel add-in "Refinitiv Eikon Datastream". The "Time Series Request" function was used to obtain data over a specific time horizon. Figure 4.2 illustrates the different inputs for the "Time Series Request". First, one or more companies are selected in the "Series" field for the query. Since the cyber news data is provided per ISIN, the same company identifier can be used for this query. For "Datatypes/Expression", the input "P" is chosen, which stands for "Price". According to the definition provided by LSEG, this represents the exchange close price displayed in the respective market currency and is adjusted for corporate actions. Next, the start and end dates for the query are defined based on the available dates in the cyber news dataset, and the frequency is set to daily. After clicking the "Submit" button, the stock prices are downloaded into an Excel sheet. This Excel sheet is then imported using the pandas package, making the data available for further analysis.

The data from the two sources differ in one important aspect. Cyber news data is generated daily, as news can occur on any day of the week. In contrast, stock prices are only available on days when stock exchanges are open, as explained in Section 2.3. Two possible solutions were considered to address this discrepancy, each based on different assumptions. The first assumption suggests that (potential) investors do not react to cyber news released on weekends or public holidays. Under this premise, such data could simply be removed from the cyber news dataset. The alternative assumption is that (potential) investors react at the earliest on the next possible trading day. For example, if news about a cyberattack is released on a Saturday, investors might buy or sell stocks, at the earliest,

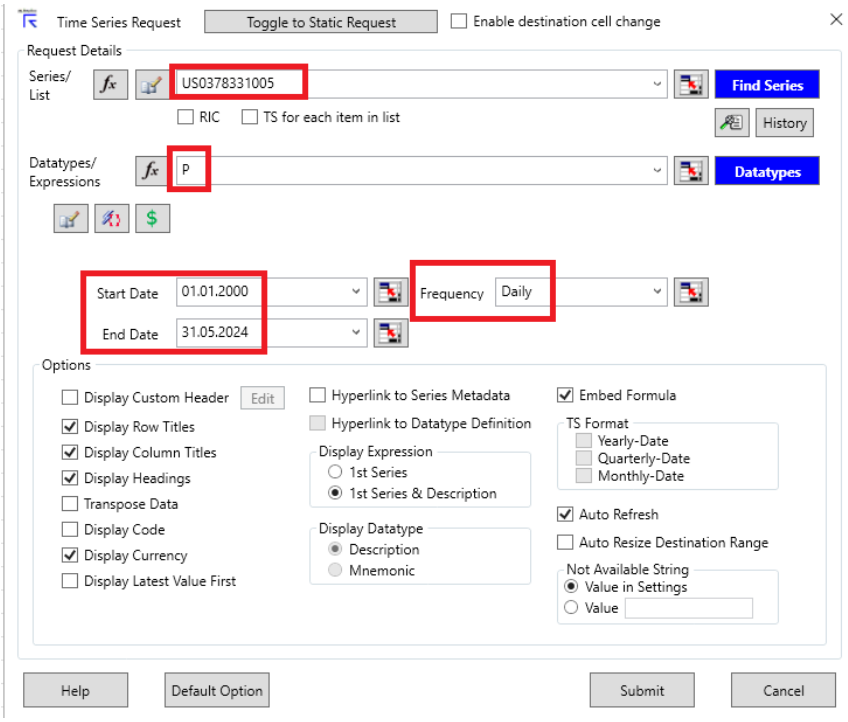


Figure 4.2: Input to Time Series Request

on the following Monday based on their assessment. Since there is often relevant news across various domains outside of stock exchange trading hours, it is likely that investors regularly check news platforms regardless of trading hours or days. Furthermore, the financial market operates globally, meaning individuals from different time zones can place stock orders at any time. These trades are then executed as soon as the relevant stock exchange reopens, as discussed in Section 2.3. Given these considerations, the second assumption was deemed more plausible.

To resolve the discrepancy between the two datasets, it was decided to use a Pandas package called "pandas_market_calendars" to identify which days are market holidays [53]. In the Cyber News dataset, data from the columns "Cyber Attack", "Cybersecurity", "Data Breach", "Data Security Management", and "Volume of News" on days without stock prices are added to the data of the next available stock trading day. The column "Perc. of Positive Sentiment" is not copied, as it represents a value between 0 and 1, and adding it would result in an artificially inflated value. Additionally, a thorough review of the data revealed that sentiment values from weekends are, in most cases, identical to those of the following trading day. Therefore, no adjustment to this column is necessary. In the stock price dataset, stock prices for holidays are simply duplicated from the previous trading day, allowing the holiday duplicates to be easily removed from the dataset.

4.2 First Scenario with Real Data

The first scenario is implemented using real cyber news data and Apple's actual stock prices. Apple is a well-known globally active designer and manufacturer of electronic

devices, including smartphones and tablets, with the iPhone contributing the largest share of its total net sales. In addition to hardware, the company offers services such as music streaming and cloud solutions and develops software and the operation system inhouse. The company's stock is listed on the Nasdaq Stock Market LLC [2].

In the initial step of this scenario, the data used for analysis is examined in more detail. The cyber news dataset covers the period from January 1, 2000, to May 31, 2024. However, there are either no data or very limited data available for the early years of this range. Figure A.1 in the appendix illustrates the development of the amount of cyber news data. The "Total Cyber News Data" column is manually created and represents the sum of the columns "Cyber Attack", "Cyber Security", "Data Breach", and "Data Security Management". The figure shows a significant increase in data starting around 2015. The reason for this rise is unclear. It could reflect an actual increase in cyber incidents, but it is also possible that the topic of cybersecurity simply gained more attention in the news. Another plausible explanation could be the increasing use of the Internet and online media, which has led to more online articles being published, resulting in a greater amount of data.

A deeper understanding of the cyber news data can be achieved through a statistical analysis, which sheds light on its structure and distribution. Figure 4.3 presents the results of this analysis.

	Volume of News	Perc. of Positive Sentiment	Cyber Attack	\
count	6142.000000	6142.000000	1323.000000	
mean	2199.169489	0.570017	13.201058	
std	4145.116283	0.053508	35.770136	
min	0.000000	0.500000	1.000000	
25%	1.000000	0.511611	2.000000	
50%	33.000000	0.574259	4.000000	
75%	1972.750000	0.625668	10.000000	
max	28999.000000	0.639062	598.000000	

	Cyber Security	Data Breach	Data Security Management
count	845.000000	849.000000	1571.000000
mean	9.339645	5.506478	16.900700
std	32.868644	23.078982	43.561235
min	1.000000	1.000000	1.000000
25%	1.000000	1.000000	2.000000
50%	3.000000	2.000000	6.000000
75%	6.000000	4.000000	16.000000
max	552.000000	441.000000	835.000000

Figure 4.3: Statistical Description of Cyber News DataFrame

The "Volume of News" and "Perc. of Positive Sentiment" columns each have 6'142 entries, indicating that, unlike the other columns, they contain no NULL values but instead zeros. Furthermore, the significantly higher mean, the distribution of the 50% and 75% quantiles, and the maximum value suggest that the majority of the "Volume of News" is not directly related to cyber columns from the dataset. In addition, the columns "Cyber Attack", "Cyber Security", and "Data Breach" exhibit relatively similar distributions. However,

the "Cyber Attack" column has the highest mean and the largest maximum value among them. The "Data Security Management" column shows slightly higher values than the previous three columns, but its distribution remains unremarkable. Finally, the "Perc. of Positive Sentiment" column reveals an interesting insight. The values in this column range between 0.5 and 0.64, indicating a relatively narrow distribution. According to the definition in Section 4.1, this suggests that the sentiment was never negative but also not strongly positive.

After completing the analysis of the cyber news data, the focus now shifts to examining Apple's stock prices in greater detail. As the cyber news dataset, the stock price dataset contains 6'142 entries, as illustrated in the statistical description presented in Figure A.3. The statistics reveal a large difference between the minimum and the maximum stock price as well as a huge standard deviation. Furthermore, the trends depicted in Figure A.2 clearly show that Apple's stock price has increased significantly over the years.

Now that the data has been thoroughly examined and its characteristics are well understood, the implementation begins. The goal of this analysis is to explore the relationship between stock prices and cybersecurity news. More specifically, the aim is to determine how cybersecurity news published today impacts stock prices in the days that follow. To establish a connection between current and future stock prices, the percentage change, known as HPR, is measured. This concept has already been explained in detail in Section 2.3. To calculate the percentage change in Apple's stock price flexibly over different time periods, a function named perfcalc, based on the HPR concept, was implemented. For the sake of simplicity, only the capital return is taken into account.

Figure 4.4 illustrates the implementation of this function. It requires three input factors: "Start_date", "Period_Days", and "ISIN". "Start_date" represents the date of the cyber news event. The second input, "Period_Days", specifies the time horizon over which the performance should be calculated. For example, if the cyber event occurred on December 1, 2023, and "Period_Days" is set to 2, the performance will be calculated between December 1, 2023, and December 3, 2023. Negative periods can also be selected to analyze whether the event had already impacted the stock price before becoming public knowledge. An essential functionality has been implemented for both the starting and ending prices. The code verifies whether the specified dates exist within the dataset. While the dataset has been cleaned to exclude weekends and stock market holidays, selecting certain periods might result in a non-existent end date. The code addresses this by incrementing the date until a valid stock price is found. For added robustness, this functionality is also applied to the starting price, ensuring accuracy when the function is executed manually without the cleaned dataset.

For example, if the starting date is Friday, December 22, 2023, and a 1-day performance calculation is requested, the code first checks whether a stock price exists on Saturday, December 23, 2023. If no price is found, it proceeds to check Sunday, December 24, 2023, and then Monday, December 25, 2023. Finally, a stock price is found on Tuesday, December 26, 2023, which is used as the next valid date for the calculation. To prevent infinite loops, the start and end dates are compared against the minimum and maximum dates in the stock price dataset. Start and end dates must always be greater than the minimum date and less than the maximum date in the dataset. The third and final input is

```python
def perfcalc(Start_date, Period_Days, ISIN):
    # Define valid date range to prevent infinite loops
    min_date = Shareprice['Date'].min()
    max_date = Shareprice['Date'].max()

    while True:
        if Start_date < min_date or Start_date > max_date:
            raise ValueError("Start_date is out of range.")
        try:
            # Try if a shareprice exists on selected Start_date
            Shareprice_Start = Shareprice.loc[Shareprice['Date'] == Start_date, ISIN].values[0]
            break
        except IndexError:
            # Move to the next day if Start_date is not found
            Start_date += pd.Timedelta(days=1)

    # Calculate End_Date based on the period
    End_Date = Start_date + pd.Timedelta(days=Period_Days)

    while True:
        if End_Date < min_date or End_Date > max_date:
            raise ValueError("End_Date is out of range.")
        try:
            # Try if a shareprice exists on selected End_Date
            Shareprice_End = Shareprice.loc[Shareprice['Date'] == End_Date, ISIN].values[0]
            break
        except IndexError:
            # Move to the next day if End_Date is not found
            End_Date += pd.Timedelta(days=1)

    # Calculate performance
    Performance = (Shareprice_End / Shareprice_Start) - 1

    return Performance
```

Figure 4.4: Implementation of the Performance Calculation Function

the company identifier ”ISIN” for which the performance calculation should be conducted. A major advantage of this function is that the result is expressed as a percentage rather than an absolute change. This facilitates comparisons between different stocks, regardless of the scale or the currency of their stock prices.

Figure 4.5 illustrates how the perfcalc function is called. It is applied to all event dates in the cyber news dataset for three different time periods: +1-day, +2-days, and +3-days. Using these inputs, the analysis seeks to examine whether an event occurring on day x has a measurable impact on the stock price in the subsequent days when compared to the stock price on day x. For each time period, a new column containing the corresponding performance values is added to the existing DataFrame.

These three time periods were chosen to find out how quickly a reaction to the cyber news might be reflected in stock prices. It could be that a reaction is visible the very next day,

```
period_days= [1,2,3]

for date in df['Date']:
    for period in period_days:
        new_column_name = 'Perf_' + str(period) + '_Days'
        df.loc[df['Date'] == date, new_column_name] = perfcalc(date, period, df.loc[df['Date'] == date, 'Company'].values[0])
```

Figure 4.5: Call of the Performance Calculation Function

or it might take longer for the news to spread among (potential) investors. Additionally, the aim is to explore whether there are differences between the event categories. For instance, (potential) investors might react faster and sell shares when hearing about a data breach compared to general cybersecurity news, where they might wait and observe how the market reacts or if further news emerges. These different time periods were selected to capture such potential differences and response patterns.

After calculating the +1-day, +2-day, and +3-day performance for the entire cyber news dataset, the final step before applying the machine learning model takes place. In this step, three different datasets are created based on distinct time periods. The idea is to investigate whether selecting different observation periods for the cyber news data affects the analysis results. As shown in Figure A.1, there were few or no cyber news data points in the early years, which might make it challenging to detect a relationship between cyber news and stock prices over the 24-year period. To investigate this, the first dataset, "df_max", includes the entire time range from January 2000 to May 2024. The second dataset aims to cover as long a time horizon as possible while focusing on the period when cyber news data became more relevant, avoiding periods with only a handful of reports. After careful analysis and testing of various thresholds, a minimum of 100 cyber news reports per day was determined as the starting point for the second dataset, named "df_2015". The third dataset "df_2021" focuses exclusively on strong signals, using a threshold of at least 600 cyber news reports per day as a starting point. Table 4.2 provides an overview of the three datasets, their respective time periods and the selection criteria.

Table 4.2: Analysis Periods for Cyber News Data

Dataset	**Time Period**	**Selection Criteria**
df_max	January 2000 - May 2024	Maximum available time period
df_2015	September 2015 - May 2024	First instance where total cyber news per day >= 100
df_2021	January 2021 - May 2024	First instance where total cyber news per day >= 600

If a relationship exists between the cyber news data and future stock prices, it should be possible to predict future performance based on the cyber news data. To test this, the three described datasets are used to train both a linear and a non-linear machine learning model. Since the task involves predicting performance as a continuous value, a regression model seems appropriate [50]. For the linear model, the ordinary least squares method provided by the open-source library scikit-learn is employed [47]. The following variables are considered:

Independent Variables:

- Volume of News, Cyber Attack, Data Security Management, Cyber Security, Data Breach, Perc. of Positive Sentiment

Dependent Variables (Targets):

- +1-day stock performance, +2-day stock performance, +3-day stock performance

Before running the regression, the three datasets need to be slightly modified. All NULL values in the dataset are replaced with 0 to maintain consistency across all columns, reflecting the condition "no data available on this day". Additionally, all independent variables are normalized using Min-Max normalization to account for the significant differences in their scales. For instance, the "Perc. of Positive Sentiment" ranges from 0 to 1, while the "Volume of News" ranges from 0 to 20,834. After preprocessing, the dataset is split into 80% training data and 20% test data, and the model is trained.

For the implementation of the non-linear model, random forest regression was chosen. Random Forest is not only known for its high accuracy but also for its ability to easily evaluate the importance of individual variables within the model [5]. For this purpose, the RandomForestRegressor functionality from the scikit-learn library was used [49]. The same normalized dataset as for the linear model served as input.

The results of the linear regression are presented in Table 4.3. Although the Mean Squared Error (MSE) is very low, the overall performance of the model is unsatisfactory. This is evident from the R^2 score. According to the scikit-learn definition, an R^2 score of 1 indicates that the model perfectly predicts the data, while a score of 0.0 means that the model always predicts the mean of the target variable, regardless of the input [48]. The R^2 scores of 0 or even slightly negative demonstrate that the linear model does not outperform a simple average prediction for the target variables. The last column, labeled "Coefficient", presents a similar outcome, as most coefficients are very close to 0. This indicates that no significant linear relationships could be identified between the independent variables and the stock performance over different time horizons.

The results of the random forest regression are summarized in Table 4.4. The MSE for this model is also close to zero, while the R^2 value is negative in all cases except for df_max and Perf_3_Days. However, the feature importance analysis reveals an interesting insight: the first and last features of the model, "Volume of News" and "Perc. of Positive Sentiment", consistently exhibit higher importance across all analyzed time periods compared to the other variables.

Table 4.3: Linear Regression Results for Scenario 1 (Apple)

Dataset	**Target**	**MSE**	**R2 Score**	**Coefficients**
df_max	Perf_1_Days	0.000534932	-0.000485856	[0.0016, 0.0043, -0.006, -0.0009, 0.0127, -0.0005]
df_max	Perf_2_Days	0.000995648	-0.000839991	[0.001, -0.0115, 0.0066, -0.0042, 0.01, -0.0004]
df_max	Perf_3_Days	0.00146265	0.000544645	[0.0016, -0.0262, 0.0042, 0.0063, 0.011, -0.0017]
df_2015	Perf_1_Days	0.000259795	0.00228197	[0.0031, 0.0054, -0.0015, -0.0022, 0.0061, -0.0029]
df_2015	Perf_2_Days	0.000542674	0.00495585	[0.0044, -0.0146, 0.0015, -0.0029, 0.0021, -0.0032]
df_2015	Perf_3_Days	0.000726985	0.00799235	[0.006, -0.0234, 0.0029, 0.0047, 0.0031, -0.0041]
df_2021	Perf_1_Days	0.000246434	-0.0199148	[-0.0028, 0.0037, -0.0122, 0.0085, 0.0162, -0.0024]
df_2021	Perf_2_Days	0.000429371	-0.00770484	[-0.001, 0.0003, -0.002, 0.0057, 0.0124, -0.0023]
df_2021	Perf_3_Days	0.000621551	-0.0185621	[-0.0066, -0.0091, -0.0003, 0.0091, 0.0127, 0.0003]

Table 4.4: Random Forest Regression Results for Scenario 1 (Apple)

Dataset	**Target**	**MSE**	**R2 Score**	**Feature Importances**
df_max	Perf_1_Days	0.000661989	-0.251589	[0.2288, 0.0378, 0.0368, 0.0184, 0.0215, 0.6567]
df_max	Perf_2_Days	0.00110376	-0.10952	[0.2219, 0.034, 0.0335, 0.02, 0.0231, 0.6674]
df_max	Perf_3_Days	0.00119366	0.0513173	[0.217, 0.0341, 0.035, 0.0193, 0.0215, 0.6731]
df_2015	Perf_1_Days	0.000300104	-0.152521	[0.2939, 0.1068, 0.1306, 0.0699, 0.0703, 0.3286]
df_2015	Perf_2_Days	0.000578632	-0.0609761	[0.2813, 0.1161, 0.1262, 0.0797, 0.0743, 0.3223]
df_2015	Perf_3_Days	0.000753196	-0.0277731	[0.2902, 0.1083, 0.1202, 0.0723, 0.0743, 0.3347]
df_2021	Perf_1_Days	0.000288614	-0.113179	[0.2353, 0.1433, 0.1567, 0.1056, 0.1066, 0.2525]
df_2021	Perf_2_Days	0.000463952	-0.088836	[0.2297, 0.1456, 0.1562, 0.1172, 0.0967, 0.2564]
df_2021	Perf_3_Days	0.000689719	-0.0958575	[0.2367, 0.1414, 0.1586, 0.1091, 0.0937, 0.2606]

4.3 Second Scenario with Synthetic Data

Apple is known for being particularly secure in the field of cybersecurity, standing out compared to its competitors. This public perception is supported by a research paper by Garg et al., which compared Apple's smartphone platform, iOS, with Android. The paper concluded that iOS is more secure than Android, primarily because Android is more vulnerable to malware attacks and security breaches. Android is, in contrast to iOS, open-source, which offers the advantage of greater customization and flexibility but also makes it more susceptible to security vulnerabilities [19]. As stated in Section 4.2, the iPhone represents the largest share of Apple's total net sales, which makes this study highly relevant to the company as a whole.

Based on the findings of this research paper, it is suggested that other companies may experience cybersecurity issues more frequently and with greater severity. As a result, there might be a stronger connection between the cyber news dataset and the stock prices of these companies compared to Apple. To test this hypothesis, Samsung Electronics Co., Ltd. (Samsung) is analyzed. Samsung is a global producer of smartphones, TVs, displays, home appliances, and a variety of connected systems and services [44]. Moreover, it uses Android rather than proprietary software for its smartphones [45]. Samsung's original shares are listed on the Korea Exchange (KRX), while other stock types are also traded on the stock exchanges in London and Luxembourg [46]. For this thesis, the stock price of the original shares listed on the KRX is used, which is published in KRW (Korean Won).

The stock prices were downloaded using the LSEG Datastream login, as previously described. Since the stock is listed on the Korea Exchange (KRX), public holidays in the Asia-Pacific region were applied for holiday adjustment. The general Asia-Pacific holidays (ASX) from the pandas_market_calendars package were used for this purpose. The Samsung stock dataset contains 6'179 entries, as shown in Figure A.5. Overall, the stock price has performed positively since 2000, although there have been some significant declines, such as in 2021. This trend is illustrated in Figure A.4.

Subsequently, the perfcalc function shown in Figure 4.4 was used to calculate stock performance for Samsung. The only difference from the invocation shown in Figure 4.5 is that the ISIN KR7005930003 was used instead of the company identifier from the cyber news dataset for Apple.

Before running the regressions, the data needs to be modified as described in the introduction of Chapter 4. In this scenario, the actual stock prices of a company, in this case, Samsung, are compared with adjusted cyber news data originally associated with Apple. Given the assumption that Samsung is more frequently affected by cyber issues, it receives greater media coverage on these events. Furthermore, these issues are believed to have a more negative impact. Based on this, the cyber news dataset is adjusted as follows:

- **More company-specific cyber news:** The columns `['Cyber Attack', 'Data Security Management', 'Cyber Security', 'Data Breach']` are squared. This ensures that days without cyber news remain unchanged, while days with cyber news are amplified.

- **Less positive sentiment:** For the column `['Perc. of Positive Sentiment']`, if the value is greater than 0.5, 0.5 is subtracted. The rationale behind this adjustment is to keep neutral days unaffected while reducing the intensity of positive sentiment. Since there are no values below 0.5, no adjustment logic needs to be implemented for this case.

After these adjustments, the modified data is split into the same three time horizons as the Apple data: df_max, df_2015, and df_2021. The same linear and non-linear regression methods are then applied, using identical dependent and independent variables.

The results of the linear regression are shown in Table 4.5. Even though the MSE values are small the R^2 Scores are close to zero or even negative across all datasets and targets. This indicates that the model has only limited explanatory power regarding the stock price predictions. In addition, the fluctuations in the coefficients provide evidence that no variable can be determined as the most important one for an accurate prediction. In comparison, the random forest regression, presented in Table 4.6, provides slightly better results. While the MSE values are also small, a slight improvement in the R^2 scores is noticeable, and there are some positive values. This indicates a marginally better performance. However, similar to the linear regression, the overall explanatory power of the random forest regression is still limited. The feature importance values show that some features contribute more consistently and again, no single variable dominates the predictions. Overall, it can be concluded that the random forest model captures some of the relationships between the cyber news data and the stock prices, but it is still not able to appropriately explain the variance in the target variables.

Table 4.5: Linear Regression Results for Scenario 2 (Samsung)

Dataset	**Target**	**MSE**	**R2 Score**	**Coefficients**
df_max	Perf_1_Days	0.000054989	-0.00004741	[-0.0016, 0.0001, -0.0, 0.0121, 0.0047, -0.0009]
df_max	Perf_2_Days	0.00012154	-0.00007845	[-0.0022, -0.0183, 0.0023, 0.0249, 0.0058, -0.0018]
df_max	Perf_3_Days	0.000216809	-0.00002589	[-0.0007, -0.0229, 0.0002, 0.0283, 0.0059, -0.0062]
df_2015	Perf_1_Days	0.000024448	-0.00010765	[-0.0015, 0.0147, 0.0098, 0.0004, 0.0132, -0.0033]
df_2015	Perf_2_Days	0.000049375	0.000218801	[-0.0013, -0.0005, -0.0007, 0.0013, 0.0132, -0.0052]
df_2015	Perf_3_Days	0.000069326	-0.00076572	[0.0018, 0.0131, -0.021, -0.0008, 0.0125, -0.0086]
df_2021	Perf_1_Days	0.000018753	-0.00013224	[-0.0008, -0.003, -0.0018, 0.0004, 0.006, -0.0]
df_2021	Perf_2_Days	0.000035027	-0.000963977	[-0.0, -0.0169, -0.0016, -0.0028, 0.0057, 0.001]
df_2021	Perf_3_Days	0.000045468	-0.00021744	[0.0011, -0.0196, -0.0047, -0.0014, 0.0047, 0.0015]

Table 4.6: Random Forest Regression Results for Scenario 2 (Samsung)

Dataset	**Target**	**MSE**	**R2 Score**	**Feature Importances**
df_max	Perf_1_Days	0.000548361	-0.0859382	[0.2389, 0.031, 0.0356, 0.0195, 0.0229, 0.6521]
df_max	Perf_2_Days	0.000797828	0.0407594	[0.2269, 0.0322, 0.0315, 0.0187, 0.0197, 0.6709]
df_max	Perf_3_Days	0.00105616	0.171833	[0.2246, 0.0319, 0.0316, 0.0187, 0.0182, 0.675]
df_2015	Perf_1_Days	0.000263266	-0.0876582	[0.3107, 0.093, 0.1276, 0.0654, 0.0754, 0.3279]
df_2015	Perf_2_Days	0.000511005	-0.0327558	[0.3101, 0.0976, 0.1182, 0.0664, 0.0767, 0.3327]
df_2015	Perf_3_Days	0.000693261	-0.00648647	[0.2979, 0.106, 0.1108, 0.0682, 0.0755, 0.3415]
df_2021	Perf_1_Days	0.000218464	-0.169712	[0.2829, 0.1217, 0.1546, 0.0987, 0.0982, 0.2439]
df_2021	Perf_2_Days	0.000391195	-0.117654	[0.2494, 0.119, 0.1661, 0.0989, 0.1054, 0.2612]
df_2021	Perf_3_Days	0.00047879	-0.0534895	[0.2467, 0.1324, 0.1591, 0.0924, 0.0955, 0.2739]

Chapter 5

Discussion and Results

Chapter 5 provides a in-depth analysis and comparison of the results from Sections 4.2 and 4.3 and it highlights the main findings and differences. It also discusses further potential use cases for cyber news data and the limitations arising from its utilization.

5.1 Scenario Comparison

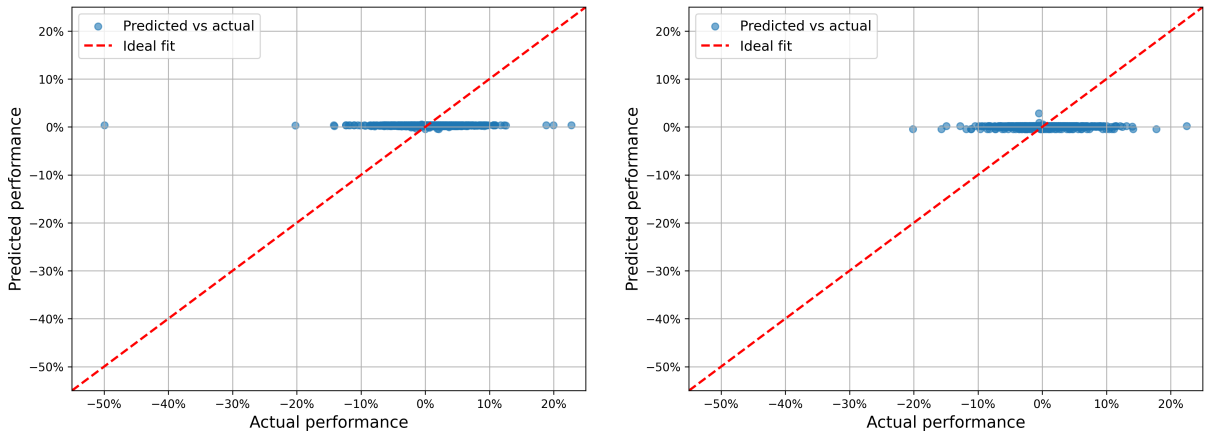
To begin with, the results of the two scenarios from the linear regression are compared. The first part of the comparison is based on the results presented in Tables 4.3 and 4.5, which list the MSE values, R^2 scores, and coefficients. While all MSE values are practically zero, it is important to note that this metric depends on the scale of the target variables. Since the target variables in this regression are percentages, small MSE values do not necessarily indicate a well-fitting model. To better assess the model's quality, the interpretation of the R^2 scores is more meaningful. The R^2 scores indicate that the model explains almost none of the variance in the target variables, as all values are close to zero. Moreover, no clear pattern emerges to indicate which time horizons or target variables perform better than others. Overall, the R^2 scores of Scenario 1 are slightly better due to fewer negative values. However, since these values are still near zero, they hold little significance. Similarly, the coefficients do not exhibit a consistent pattern, either in their importance to the model or in whether they indicate a positive or negative relationship with the target variable.

In the second part of the scenario comparison for the linear regression, the predicted performance values are compared with the actual performance values. This comparison is illustrated in Figures 5.1 (a) and (b). Both figures are based on the dataset with the maximum time horizon (df_max) and the target variable representing the +3-day stock performance (Perf_3_Days). This combination of time horizon and target variable yielded the best R^2 score across the entire comparison. As mentioned, the differences in R^2 scores across various combinations of datasets and targets are minimal, meaning they are barely distinguishable in the visualization of predicted vs. actual performance. Therefore, the same combination was chosen for better comparability with the random forest regression.

The two figures feature a red dashed line labeled ”Ideal fit”, representing the line on which perfectly predicted data points would lie. If a model achieved an R^2 score of 1, all predicted performance values would align perfectly with this line.

Figure 5.1a depicts the comparison for Scenario 1 using the real cyber news and stock data of Apple. Here, the predicted performance remains close to 0% regardless of the actual performance. Even for outliers with a +3-day performance of -50%, -20%, and +20%, the model predicts a performance of 0%. The outlier with approximately -50% performance occurred around September 29, 2000. However, during this period, the columns ”Volume of News,” ”Cyber Attack,” ”Data Security Management,” ”Cyber Security,” and ”Data Breach” are zero, while the ”Perc. of Positive Sentiment” column shows a neutral value of 0.5. Even if this drop was not caused by cyber news, it is implausible for the ”Volume of News” to be zero and for sentiment to remain neutral during such a significant decline. This raises questions about the overall quality of the data, particularly the quality of the historical cyber news data. For another outlier, the -20% performance decline around September 25, 2008, some data for ”Volume of News” and certain cyber news categories do exist, and the ”Perc. of Positive Sentiment” is not neutral. Despite the availability of data for this event, the linear regression was unable to predict the performance drop.

Figure 5.1b visualizes the same comparison between predicted and actual performance for the second scenario, which uses the modified data for Samsung. Similarly, the predicted performance values form a horizontal line at approximately 0%. In Samsung's +3-day performance data, there are also a few outlier values where the predictions deviate slightly from 0%, yet remain very close to it.



(a) Scenario 1 (Apple)

(b) Scenario 2 (Samsung)

Figure 5.1: Comparison Between the Predicted and the Actual Performance for the Linear Regression

In summary, the linear model does not provide satisfactory predictions for either Scenario 1 or Scenario 2. This indicates that no linear relationship between the available cyber news data and the stock prices of Apple and Samsung can be identified in the existing dataset. However, it is important to note that only a limited cyber news dataset was available for

the analysis. The potential limitations arising from this constraint are discussed in more detail in Section 5.2.

The analysis now turns to the results of the random forest regression. As with the linear model, the results are first compared using different metrics, followed by a visual comparison. The metrics are listed in Tables 4.4 and 4.6. While the MSE values for these regressions are also minimal, the same explanation as for the linear model applies, which limits the interpretability of this metric. The R^2 scores, however, deviate further from zero, tending to be either more strongly positive or negative. Furthermore, the only dataset that achieves a positive R^2 score across both scenarios is df_max. This dataset delivers the highest R^2 score for both Scenario 1 and Scenario 2 when using the Perf_3_Days target variable. Interestingly, this results in the highest R^2 score across all scenarios, models, datasets, and target variables.

This raises the question of why this particular combination resulted in the best R^2 score. To explore a possible answer, the two components will be analyzed separately. The improved performance of df_max might be attributed to the fact that it contains the largest number of data points, and random forest models are generally designed to perform well with large datasets. However, this theory is contradicted by the exceptionally poor R^2 score of the df_max and Perf_1_Days model of -0.25 in Scenario 1. To further investigate this, more cyber news data from other companies would be required. Regarding Perf_3_Days, one possibility is that the effects from the cyber news dataset take more than one or two days to influence stock prices. However, there could be other explanations as well. For instance, the +3-day performance might combine the effects of multiple days, creating a stronger overall impact on the stock price in relation to the news data. To investigate this further, additional time horizons could be examined in future studies. Another plausible explanation is that data noise is generally reduced over longer time periods, as random fluctuations tend to balance out. In conclusion, the improved performance of the df_max and Perf_3_Days combination may result from the larger dataset size and cumulative effects over longer time horizons. Nonetheless, inconsistencies with other combinations and the potential influence of noise reduction suggest further investigation is needed.

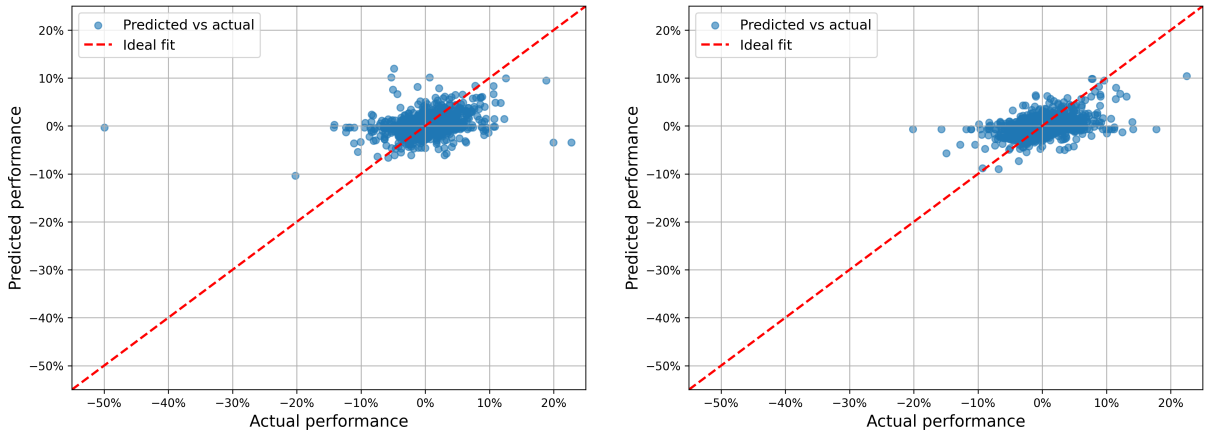
There are also similarities between the two scenarios regarding feature importance. In both cases, the variables "Volume of News" and "Perc. of Positive Sentiment" stand out due to their higher importance, with comparable magnitudes. Nevertheless, these two variables do not have a direct connection to cyber news. While cyber news is included in the total news volume, this metric may also encompass news from other event categories, such as a company's financial situation. Similarly, "Perc. of Positive Sentiment", as defined, is not exclusively connected to cyber news. This observation suggests that news data, in general, may have an impact on stock prices, but the specific cyber news included in the dataset does not necessarily play a significant role.

Visualizations were also created for the random forest regression to compare the predicted performance with the actual performance. For Figures 5.2a and 5.2b, the dataset df_max and the target variable Perf_3_Days were again used.

In Figure 5.2a, the values show some scatter and several points deviate notably from the "Ideal fit" line. Outliers are also apparent in this case. The -50% decline could not be

predicted, likely due to missing data. For the -20% decline, cyber news data is available as previously mentioned. In this case, the model predicted a performance of approximately -10%. Although this prediction is not accurate, it does at least capture the correct trend.

The results for Scenario 2 are shown in Figure 5.2b. The scatterplot appears somewhat more compact with points positioned closer to the "Ideal fit" line compared to the Figure 5.2a. In addition, there seem to be fewer outliers. In summary, this indicates a modest improvement in predictive performance.



(a) Scenario 1 (Apple)

(b) Scenario 2 (Samsung)

Figure 5.2: Comparison Between the Predicted and the Actual Performance for the Random Forest Regression

When comparing the results of the linear regression and the random forest regression, several differences become apparent. For instance, when examining Figures 5.1 and 5.2, it is immediately noticeable that the predicted performance values no longer form a horizontal line near zero but instead create a scattered point cloud. Additionally, the overall best R^2 value was achieved with the random forest regression. However, it should also be noted that the R^2 values for the random forest regression are generally further from zero compared to the linear model, both in the positive and negative directions. This indicates that there are combinations of time periods and target variables where the random forest regression performs worse than the linear model. The consistently low R^2 scores overall may point to factors such as insufficient data for the analysis, the presence of data inconsistencies, or the absence of key variables critical for the regressions.

Regarding the outliers, an improvement was observed in the random forest model compared to the linear model. The substantial decline of -20%, for which cyber news data is available, is more accurately captured by the random forest model.

An additional point of consideration is whether the modifications made in Scenario 2 had a noticeable effect. While Scenario 2 achieved the best R^2 score, the differences between the two scenarios remain modest overall. This suggests that the modifications likely had only a limited impact.

In conclusion, the random forest regression seems to slightly better capture the relationship between the cyber news data and stock prices. However, it is important to emphasize again that only the cyber news data of a single company was available for the entire analysis. A detailed discussion of the limitations can be found in Section 5.2.

5.2 Opportunities and Limitations

As highlighted in Chapter 3, leveraging cyber news data offers a novel approach to quantifying cyber risks. These non-anonymized, almost real-time datasets are updated daily and available at the company level. This provides a notable advantage compared to traditional data sources like risk reports, which are often anonymized and infrequently published.

Cyber news data could also be useful in other areas, such as corporate cyber risk management. For instance, these data could serve as an early warning system for cyber threats in general or in connection with specific companies. If a cybersecurity manager notices an increase in reports about data breaches, the corresponding controls and protective measures in that area could be intensified. Moreover, cyber news data could find applications in the financial sector, particularly for trading strategies. An algorithm-based trading strategy could analyze the volume of news in specific categories and determine whether they suggest a positive signal (to buy stocks) or a negative signal (to sell stocks). Another promising application is in customer decision-making. Before customers purchase, for example, a smartphone, they could compare cyber news data to evaluate the cybersecurity standards of different providers. This concept extends beyond personal electronic devices, as many products today are interconnected. As a result, assessing cybersecurity levels could also be useful before buying cars, refrigerators, or other connected appliances.

In addition to the opportunities that the use of cyber news data presents, it also entails certain risks. The data could, for instance, be manipulated by individuals with malicious intent. For instance, fake news articles could be created to generate specific signals, which allow perpetrators to gain a personal advantage. Another potential risk lies in unauthorized modifications to the algorithms that analyze cyber data and generate the datasets. Even minor changes to these algorithms could result in events from certain companies being excluded from the dataset. Companies could use such manipulations to present their cybersecurity situation in a more favorable light than it actually is.

As outlined in Chapter 4, this thesis was limited to use a restricted dataset of cyber news data. This limitation introduces potential weaknesses. First, the exact process by which the data was collected, as well as its precise interpretation, remains unclear. Since the algorithm used to generate the cyber news dataset has not been disclosed, there is some uncertainty about which online media sources were analyzed, how events have been verified before they were included in the dataset, and how they were categorized into types such as ”Data Security Management” or ”Data Breach”. Additionally, the category ”Total Volume of News” lacks transparency regarding which event types are included. There are substantial discrepancies between the total volume of news and the sum of cyber event news, suggesting that many other event categories outside the cyber domain are also represented. A more comprehensive understanding of the relationship between news data

and stock prices might be achieved if all event types were included in the dataset. The metric "Perc. of Positive Sentiment" introduces as well some ambiguity. According to the definition, it represents the proportion of positive mentions relative to total mentions but the criteria to classify mentions as positive or negative are undefined.

The dataset includes data from only one company, namely Apple. With more data, it would be possible to investigate whether other companies also experienced an increase in cyber news data around 2015. This would help to determine whether the surge was specific to Apple, perhaps due to increased media attention, or whether it reflected a general rise in cyber news coverage. Furthermore, all sentiment scores in the current dataset are at least neutral. It would be valuable to explore the impact of negative sentiment data, particularly in the context of quantifying cyber risks. To analyze the consequences of a cyberattack, which is inherently associated with negative sentiment, it becomes essential to understand how such events affect companies. It would also be interesting to evaluate a company whose stock performance has not been as strong as Apple's. It is possible that any negative impacts of cyber news on Apple were neutralised by positive news from other areas such as rising profits. Moreover, a comparison across different industries could provide insights into patterns related to a companies' business. For example, stocks from other sectors, such as those less dependent on online business, might react differently or with a different lag to cyber news data.

Chapter 6

Summary and Conclusions

At the beginning of the thesis, the steady rise in cyber threats over the last years is shown, which highlights the growing importance for organizations of all sizes, especially SMEs. Subsequently, RCVaR is analyzed in depth. The analysis reveals that while it already provides plausible predictions of cyber risk losses, its accuracy and robustness could be further enhanced with the inclusion of additional data. Moreover, the Efficient Market Hypothesis (EMH) is introduced and its potential to improve RCVaR is highlighted. It shows that information about cyber events could be reflected in stock prices, and changes in these prices might offer insights into the distribution and magnitude of losses resulting from cyber incidents.

The examination of cyber risk frameworks revealed that many different approaches exist and each offers its own strengths and limitations. Among them, RCVaR stands out through its forward-looking perspective, broad applicability, and ease of use and understanding. Moreover, no framework was identified that already utilizes cyber news data. For practical application, a wide variety of tools is available for companies to manage and quantify their cyber risks. However, no all-in-one solution was found that addresses diverse use cases and delivers outputs that are clear, transparent, and easy to interpret.

During the implementation process, it became clear that only a limited cyber news dataset could be used for this thesis. As a result, the approach was adjusted, and two different scenarios were created to study the relationship between cyber news data and stock prices. In the first scenario, real cyber news data and stock prices are used. In the second scenario, the cyber news data was modified to reflect a negative cyber situation. For each scenario, both a linear regression and a random forest regression were implemented.

The random forest regression achieved a slightly better performance compared to the linear regression. This result belongs to the period from 2000 to 2024, which includes the impact of cyber events on stock prices after three days. However, the interpretation of the results requires careful consideration of the following points. First, the results were inconsistent across different time periods and target variables, and generally weak in their statistical significance. Second, the absence of news data and neutral sentiment data during extreme outliers raises questions about the reliability and completeness of the cyber news dataset. Third, the finding that the two most relevant features identified

by the random forest regression were not directly related to cyber news but to general news points out that a more detailed investigation of the overall influence of news data on stock prices is needed. Fourth, the overall weak connection between the datasets indicates that key variables which influence stock prices are missing in the regressions. This further limits the robustness of the analysis.

These findings suggest that the analysis should be expanded and a larger and more diverse dataset should be used. This includes data from companies of different sizes, industries and various levels of corporate success and cybersecurity maturity. Additionally, the dataset should contain more different event categories, not just cyber news. This enables a broader understanding of the overall impact of news on stock prices and plays a key role to precisely assess the impact of cyber news.

Furthermore, it is essential to gain a deeper understanding of the methodology which was used to construct the dataset. This might help to better understand the results and to provide insights into questions such as why certain data, particularly regarding outliers, is included or excluded from the dataset.

Given the limited availability of data, the findings lack sufficient robustness to meaningfully contribute to the improvement of RCVaR. As a result, incorporating them is beyond the scope of this thesis. However, the insights gained, along with the developed code for data loading, cleaning, performance calculation, and regression analyses, can be reused. Once the complete cyber news datasets become available, this groundwork will enable future investigations to be conducted efficiently.

6.1 Future Work

The primary and most critical future work includes the application of the complete cyber news dataset. Once cyber news data from multiple companies is available, both linear and random forest regression analyses can be carried out. Additionally, the analysis can be extended to include other target variables, including negative time horizons, to investigate whether cyber news data already influences stock prices before becoming part of the cyber news dataset. Furthermore, additional independent variables could be incorporated into the model, such as company size or the performance of a relevant benchmark, once the full set of companies with available cyber news data is known. Another potential area of investigation is parameter tuning for the random forest model. For example, the number of trees could be adjusted to test if this substantially and consistently improves the results. Finally, other non-linear machine learning approaches could be applied to evaluate if they outperform the random forest model. This could provide a better understanding of the optimal approaches to analyze how stock prices react to cyber news.

When the exact methodology used to create the cyber news dataset becomes clear, it could be adjusted or extended. For instance, the threshold for including news items in the dataset could be raised or lowered to find out if this has an impact on the analysis. Furthermore, additional cyber news categories, such as ”Supply Chain Attacks” could also be introduced.

The assumption that RCVaR could be further improved by making use of cyber news data is still valid. This thesis contributes to exploring this concept. However, future studies with access to more comprehensive data should continue to investigate this assumption to reach a conclusive outcome.

Bibliography

[1] M. S. Abu, S. R. Selamat, A. Ariffin, and R. Yusof. Cyber Threat Intelligence - Issue and Challenges. *Indonesian Journal of Electrical Engineering and Computer Science*, 10(1):371–379, 2018. Accessible online: `http://doi.org/10.11591/ijeecs.v10.i1.pp371-379`, Last accessed Jan. 2025.

[2] Apple Inc. Form 10-K - For the Fiscal Year Ended September 28, 2024, 2024. Accessible online: `https://s2.q4cdn.com/470004039/files/doc_earnings/2024/q4/filing/10-Q4-2024-As-Filed.pdf`, Last accessed Jan. 2025.

[3] AttackIQ. PREACT Security Optimization Program, 2024. Accessible online: `https://www.attackiq.com/solutions/preact-security-optimization/`, Last accessed Oct. 2024.

[4] C. Biener, M. Eling, and Wirfs J. H. Insurability of Cyber Risk: An Empirical Analysis. *Geneva Papers on Risk and Insurance*, 40(1), 2015. Accessible online: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2577286`, Last accessed Sep. 2024.

[5] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. Accessible online: `https://doi.org/10.1023/A:1010933404324`, Last accessed Jan. 2025.

[6] Cybersecurity Ventures. 2023 official cybercrime report, 2024. Accessible online: `https://www.esentire.com/resources/library/2023-official-cybercrime-report`.

[7] M. Ekstedt, Z. Afzal, P. Mukherjee, S. Hacks, and R. Lagerström. Yet another cybersecurity risk assessment framework. *International Journal of Information Security*, 22(6):1713–1729, 2023. Accessible online: `https://doi.org/10.1007/s10207-023-00713-y`, Last accessed Sep. 2024.

[8] A. Erola, I. Agrafiotis, J. R. C. Nurse, L. Axon, M. Goldsmith, and S. Creese. A system to calculate Cyber Value-at-Risk. *Computers & Security*, 113:102545, 2022. Accessible online: `https://www.sciencedirect.com/science/article/pii/S0167404821003692`, Last accessed Sep. 2024.

[9] FAIR Institute. FAIR: A Methodology for Quantifying and Managing Risk in Any Organization, 2024. Accessible online: `https://www.fairinstitute.org/what-is-fair`, Last accessed Sep. 2024.

[10] E. F. Fama. Random Walks in Stock Market Prices. *Financial Analysts Journal*, 21(5):55–59, 1965.

[11] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970. JSTOR.

[12] Federal Bureau of Investigation. Internet Crime Report 2023, 2023. Accessible online: `https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf`, Last accessed Sep. 2024.

[13] M. F. Franco. CyberTEA: a Technical and Economic Approach for Cybersecurity Planning and Investment. PhD Thesis, University of Zurich, 2023. Accessible online: `https://figueredofranco.com/static/files/PhD-M-Franco.pdf`, Last accessed Sep. 2024.

[14] M. F. Franco, F. Künzler, J. von der Assen, C. Feng, and B. Stiller. RCVaR: An economic approach to estimate cyberattacks costs using data from industry reports. *Computers & Security*, 139:103737, 2024. Accessible online: `https://www.sciencedirect.com/science/article/pii/S0167404824000385`, Last accessed Sep. 2024.

[15] M. F. Franco, A. R. Mullick, and S. Jha. QBER: Quantifying Cyber Risks for Strategic Decisions. 2024. Accessible online: `https://arxiv.org/abs/2405.03513`, Last accessed Sep. 2024.

[16] J. Freund and J. Jones. Chapter 3 - The FAIR Risk Ontology. In *Measuring and Managing Information Risk*, pages 25–41. Butterworth-Heinemann, Boston, 2015. Accessible online: `https://www.sciencedirect.com/science/article/pii/B9780124202313000038`, Last accessed Jan. 2025.

[17] R. Gafni and T. Pavel. The invisible hole of information on smb's cybersecurity. *Online Journal of Applied Knowledge Management (OJAKM)*, 7(1):14–26, 2019. Accessible online: `https://www.iiakm.org/ojakm/articles/2019/OJAKM_Volume7_1pp14-26.php`, Last accessed Jan. 2025.

[18] N. Gandal, M. H. Riordan, and S. Bublil. A new approach to quantifying, reducing and insuring cyber risk: Preliminary analysis and proposal for further research. 2020. Accessible online: `https://ssrn.com/abstract=3548380`, Last accessed Sep. 2024.

[19] S. Garg and N. Baliyan. Comparative analysis of Android and iOS from security viewpoint. *Computer Science Review*, 40:100372, 2021. Accessible online: `https://www.sciencedirect.com/science/article/pii/S1574013721000125?ref=cra_js_challenge&fr=RR-1`, Last accessed Dec. 2024.

[20] M. Grant. Trading Hours for the World's Major Stock Exchanges, 2024. Accessible online: `https://www.investopedia.com/ask/answers/040115/when-do-stock-market-exchanges-close.asp`, Last accessed Dec. 2024.

[21] IBM. What is cyber risk management?, 2023. Accessible online: `https://www.ibm.com/topics/cyber-risk-management`, Last accessed Oct. 2024.

[22] Internet Crime Complaint Center (IC3). Homepage, no date. Accessible online: https://www.ic3.gov/, Last accessed Sep. 2024.

[23] ISO. Iso/iec 27001:2008, 2008. Accessible online: https://www.iso.org/standard/42107.html, Last accessed Sep. 2024.

[24] ISO. Iso/iec 27005:2022, 2022. Accessible online: https://www.iso.org/standard/80585.html, Last accessed Sep. 2024.

[25] H. Jiang, N. Khanna, Q. Yang, and J. Zhou. The Cyber Risk Premium. *Management Science*, 70(12):8791–8817, 2024. Accessible online: https://doi.org/10.1287/mnsc.2022.02056, Last accessed Sep. 2024.

[26] M. Karyda. Cyber Risk Quantification. In S. Jajodia, P. Samarati, and M. Yung, editors, *Encyclopedia of Cryptography, Security and Privacy*, pages 1–3. Springer Berlin Heidelberg, Berlin, Heidelberg, 2019.

[27] H. Kavak, J.J. Padilla, D. Vernon-Bido, S.Y. Diallo, R. Gore, and S. Shetty. Simulation for cybersecurity: state of the art and future directions. *Journal of Cybersecurity*, 7(1), 2021. Accessible online: https://academic.oup.com/cybersecurity/article/7/1/tyab005/6170701?login=false, Last accessed Sep. 2024.

[28] kovrr. CRQ: The Key to Understanding and Managing Cyber Risk, 2023. Accessible online: https://www.kovrr.com/cyber-risk-quantification, Last accessed Sep. 2024.

[29] F. Künzler. Real Cyber Value at Risk: An Approach to Estimate Economic Impacts of Cyberattacks on Businesses. Master thesis, University of Zurich, 2023. Accessible online: https://www.zora.uzh.ch/id/eprint/255756/1/MA_F_Kunzler.pdf, Last accessed Sep. 2024.

[30] J. Lanz. The Updated NIST Cybersecurity Framework. *The CPA Journal*, 94(5/6):70–72, 2024. Accessible online: https://www.proquest.com/openview/cf718e5b24f6e05dea2c90de6fd9448c/1?cbl=41798&pq-origsite=gscholar&parentSessionId=lir3ZJW5q%2FvQJyu0Gn47vnpFbBSRYsXjFcCs4egtxw0%3D, Last accessed Sep. 2024.

[31] C. Lee and A. C. Lee. Terms and Essays. In C. Lee and A. C. Lee, editors, *Encyclopedia of Finance*, pages 3–507. Springer, Cham, 2022. Accessible online: https://doi.org/10.1007/978-3-030-91231-4_1, Last accessed Jan. 2025.

[32] LSEG. About LSEG. no date. Accessible online: https://www.lseg.com/en/about-us, Last accessed Jan. 2025.

[33] E. Mondello. Rendite, Risiko und Markteffizienz. In *Finance*. Springer Gabler, Wiesbaden, 2017. Accessible online: https://doi.org/10.1007/978-3-658-13199-9_2, Last accessed Jan. 2025.

[34] National Institute of Standards and Technology. Managing Information Security Risk - Organization, Mission, and Information System View , 2011. Accessible online: https://dl.acm.org/doi/pdf/10.5555/2206253, Last accessed Nov. 2024.

[35] National Institute of Standards and Technology. Framework for Improving Critical Infrastructure Cybersecurity - Version 1.0, 2014. Accessible online: https://www.nist.gov/system/files/documents/cyberframework/cybersecurity-framework-021214.pdf, Last accessed Sep. 2024.

[36] National Institute of Standards and Technology. Framework for Improving Critical Infrastructure Cybersecurity - Version 1.1, 2018. Accessible online: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf, Last accessed Sep. 2024.

[37] National Institute of Standards and Technology. The NIST Cybersecurity Framework (CSF) 2.0, 2024. Accessible online: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf, Last accessed Sep. 2024.

[38] A. Orlando. Cyber Risk Quantification: Investigating the Role of Cyber Value at Risk. *Risks*, 9(10):184, 2021. Accessible online: https://www.mdpi.com/2227-9091/9/10/184, Last accessed Sep. 2024.

[39] P. Radanliev, R. Mantilla Montalvo, Nicolescu R. Cannady, S., D. De Roure, J. R. C. Nurse, and M. Huth. Cyber Security Framework for the Internet-of-Things in Industry 4.0. March 2019. Accessible online: https://doi.org/10.20944/preprints201903.0111.v1, Last accessed Sep. 2024.

[40] A. Raghavan and A. Thomas. Deloitte review: quantifying risk. 2016. Accessible online: https://www2.deloitte.com/us/en/insights/deloitte-review/issue-19/quantifying-risk-lessons-from-financial-services-industry.html, Last accessed Sep. 2024.

[41] RiskLens. RiskLens Enterprise, 2024. Accessible online: https://www.risklens.com/products-services/platform, Last accessed Sep. 2024.

[42] S. Saeed, S. A. Altamimi, N. A. Alkayyal, E. Alshehri, and D. A. Alabbad. Digital Transformation and Cybersecurity Challenges for Businesses Resilience: Issues and Recommendations. *Sensors*, 23(15):6666, 2023. Accessible online: https://www.mdpi.com/1424-8220/23/15/6666, Last accessed Jan. 2025.

[43] J. Saleem, B. Adebisi, R. Ande, and M. Hammoudeh. A state of the art survey - Impact of cyber attacks on SME's. In *Proceedings of the International Conference on Future Networks and Distributed Systems*, ICFNDS '17, page 52, New York, NY, USA, 2017. Association for Computing Machinery. Accessible online: https://dl.acm.org/doi/abs/10.1145/3102304.3109812, Last accessed Jan. 2025.

[44] Samsung. Consolidated financial statements - december 31, 2023 and 2022, 2024. Accessible online: https://images.samsung.com/is/content/samsung/assets/global/ir/docs/2023_con_quarter04_all.pdf, Last accessed Jan. 2025.

[45] Samsung. How can I check what version of Android I have on my device?, 2024. Accessible online: https://www.samsung.com/uk/support/mobile-devices/how-can-i-check-what-version-of-android-i-have-on-my-device/, Last accessed Dec. 2024.

[46] Samsung. Listing information, no date. Accessible online: `https://www.samsung.com/global/ir/stock-information/listing-Info/`, Last accessed Dec. 2024.

[47] scikit-learn developers (BSD License). LinearRegression, 2024. Accessible online: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression`, Last accessed Dec. 2024.

[48] scikit-learn developers (BSD License). R2 Score, 2024. Accessible online: `https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.r2_score.html`, Last accessed Dec. 2024.

[49] scikit-learn developers (BSD License). RandomForestRegressor, 2024. Accessible online: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html`, Last accessed Dec. 2024.

[50] scikit-learn developers (BSD License). scikit-learn - Machine Learning in Python, no date. Accessible online: `https://scikit-learn.org/1.5/index.html`, Last accessed Dec. 2024.

[51] Securities and Exchange Commission. How to Read a 10-K, 2011. Accessible online: `https://www.sec.gov/answers/reada10k.htm`, Last accessed Jan. 2025.

[52] M. H. U. Sharif and M. A. Mohammed. A literature review of financial losses statistics for cyber security and future trend. *World Journal of Advanced Research and Reviews*, 15(1):138–156, 2022. Accessible online: `https://doi.org/10.30574/wjarr.2022.15.1.0573`, Last accessed Sep. 2024.

[53] R. Sheftel. pandas_market_calendars, 2016. Accessible online: `https://pandas-market-calendars.readthedocs.io/en/latest/`, Last accessed Dec. 2024.

[54] SOCRadar. Extended Threat Intelligence, 2024. Accessible online: `https://socradar.io/`, Last accessed Oct. 2024.

[55] A. Sukumar, H. Amoozad Mahdiraji, and V. Jafari-Sadeghi. Cyber risk assessment in small and medium-sized enterprises: A multilevel decision-making approach for small e-tailors. *Risk Analysis*, 43(10):2082–2098, 2023. Accessible online: `https://onlinelibrary.wiley.com/doi/abs/10.1111/risa.14092`, Last accessed Sep. 2024.

[56] ThreatConnect. Risk Quantifier (RQ), 2024. Accessible online: `https://threatconnect.com/risk-quantifier/#assess-cyber-risks`, Last accessed Oct. 2024.

[57] University of Zurich. Special databases. 2023. Accessible online: `https://www.ub.uzh.ch/en/unterstuetzung-erhalten/fachliche-unterstuetzung/wirtschaftswissenschaften/spezialdatenbanken.html`, Last accessed Dec. 2024.

[58] G. Wangen, C. Hallstensen, and E. Snekkenes. A framework for estimating information security risk assessment method completeness. *International Journal of Information Security*, 17:681–699, 2018. Accessible online: `https://doi.org/10.1007/s10207-017-0382-0`, Last accessed Sep. 2024.

[59] World Economic Forum (WEF). Partnering for Cyber Resilience Towards the Quantification of Cyber Threats. 2015. Accessible online: https://www3.weforum.org/docs/WEFUSA_QuantificationofCyberThreats_Report2015.pdf, Last accessed Sep. 2024.

[60] World Economic Forum (WEF). Global Cybersecurity Outlook 2024: Insight Report, 2024. Accessible online: https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2024.pdf, Last accessed Jan. 2025.

[61] X-Analytics. Align cyber risk management with business objectives, no date. Accessible online: https://www.x-analytics.com/, Last accessed Oct. 2024.

[62] Zeron. Dashboard, 2024. Accessible online: https://docs.zeron.one/docs/Dashboard/Dashboard-Intro/, Last accessed Oct. 2024.

[63] Zeron. Empowering Decisions By Quantifying Cyber Risks, 2024. Accessible online: https://zeron.one/, Last accessed Sep. 2024.

Abbreviations

SMEs	Small and middle-sized enterprises
CVaR	Cyber Value at Risk
VaR	Value at Risk
RCVaR	Real Cyber Value at Risk
SWARA	Step-wise Weight Assessment Ratio Analysis
BWM	Best-Worst Method
DFDs	Data flow diagrams
SaaS	Software as a Service
EMH	Efficient Market Hypothesis
CAPM	Capital Asset Pricing Model
ML	Machine Learning
HPR	Holding Period Return
MSE	Mean Squared Error

List of Figures

List of Tables

Appendix A

Statistical Analysis and Plots

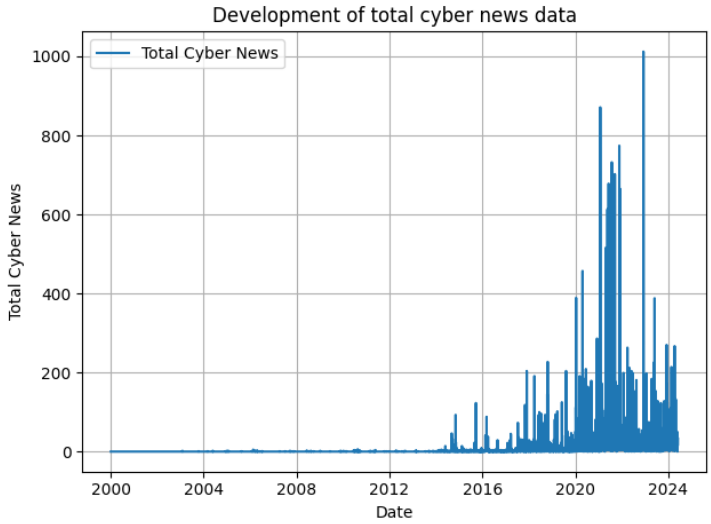


Figure A.1: Development of Total Cyber News Data

Figure A.2: Development of Apple Stock Price

count	6142.000000
mean	38.328712
std	53.972342
min	0.234300
25%	2.271625
50%	15.299650
75%	43.220625
max	198.110000

Figure A.3: Statistical Description of Apple Stock Price

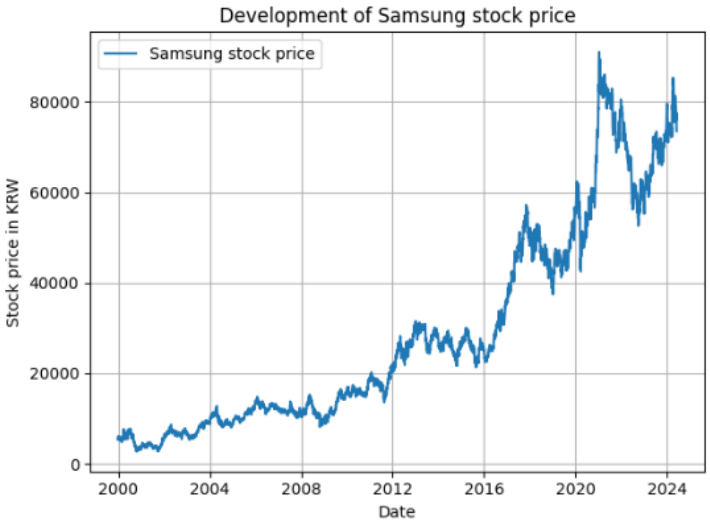


Figure A.4: Development of Samsung Stock Price

count	6179.000000
mean	28675.436645
std	22550.109517
min	2730.000000
25%	10900.000000
50%	22240.000000
75%	45980.005000
max	91000.000000

Figure A.5: Statistical Description of Samsung Stock Price