The Application of Survival Analysis in Stock Market – A Web Interface Approach

Jiale Chen

Master of Science

School of Informatics

University of Edinburgh

2023

Abstract

This project investigates the application of survival analysis techniques to model risks in the stock market. Survival analysis refers to a collection of statistical methods for analyzing the expected duration until the occurrence of an event of interest. While these techniques have proven efficacy across myriad domains, their potential in the financial realm remains relatively unexplored. This research aims to address this gap through the development of an interactive web interface tailored specifically for stock market analysis.

The study has multiple key objectives: (1) to implement robust data collection and preprocessing pipelines; (2) to build customized survival analysis models incorporating market conditions and financial indicators; (3) to design an intuitive user interface to facilitate interactive experimentation; and (4) to provide comprehensive visualizations to elucidate model outputs. This project implements a diverse array of models spanning parametric, semi-parametric, and non-parametric categories. The web interface is powered by Python libraries including Pandas, Plotly, Lifelines, and Streamlit.

This platform demonstrates the value of survival analysis in understanding stock market behaviours such as delisting risks and Value-at-Risk threshold breaches. The practical tool developed empowers users across skill levels to conduct sophisticated analytics and derive actionable insights. While limitations exist regarding model scope and data biases, the research underscores the significant potential of integrating survival analysis within financial market contexts. Avenues for future work include expanding the model horizon, incorporating broader data, and enhancing real-time analytical capabilities.

Overall, this project makes both theoretical and practical contributions, introducing survival analysis as a novel perspective for stock market analysis while also developing an accessible web-based tool to fulfil a clear industry need for robust risk assessment.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Jiale Chen*)

Acknowledgements

In the arduous journey that is the creation of an MSc dissertation, there are many who play pivotal roles in ensuring its success, and it is with profound gratitude that I acknowledge them here.

First and foremost, my deepest appreciation goes to Dr Felipe Costa Sperb, my supervisor. His vast expertise, consistent guidance, and unwavering support made the complexities of this project more navigable. His commitment to the highest standards inspired and challenged me in equal measure.

I am equally thankful to Dr Tiejun Ma. The assistance he provided at various stages of the development process was invaluable. His constructive feedback and insights greatly enriched the quality of this dissertation.

On a personal note, my gratitude extends to my girlfriend, Zhihan Yu. Her love, patience, and unwavering support were the pillars that kept me going through the ups and downs of this academic venture. She has been a beacon of hope and encouragement.

Lastly, I cannot forget the bedrock of my life – my family. Their constant belief in my potential, their love, and the sacrifices they've made have propelled me to this milestone. Their trust and encouragement served as my guiding light, and I owe them more than words can convey.

To all of you, my heartfelt thanks. This accomplishment is not just my own but a testament to the collective effort of all who believed in me.

Table of Contents

Chapter 1

Introduction

1.1 Background and Motivation

In the past several decades, the importance of risk management in stock market operations has risen to the forefront, largely driven by catastrophic financial events that highlighted the dire consequences of its oversight[1]. One of the most significant events was the 2008 financial crisis[2]. During this period, the collapse of major financial institutions like Lehman Brothers sent shockwaves throughout the global economy, underlining the critical importance of effective risk management to mitigate extreme losses.

Within the intricate ecosystem of the stock market, which includes listed companies, individual and institutional investors, exchanges, and other financial entities, the stakes are high[3]. For these stakeholders, vigilant risk management isn't just an academic concept but a necessity for preserving capital and ensuring the smooth functioning of financial markets. Historically, to gauge and respond to these risks, many experts have turned to statistical models. One of the most prevalent methodologies has been using the historical distribution of returns to estimate potential losses, a method known as Value at Risk (VaR)[4]. By leveraging such models, investors and institutions can better understand potential downside scenarios and make informed decisions that align with their risk tolerance.

The role of statistical models, such as VaR, goes beyond mere risk assessment[5]. It aids in portfolio optimization, capital allocation, and even regulatory compliance. As stock markets continue to evolve and become more complex, the need for advanced and precise risk management tools will only grow, highlighting the pivotal role of quantitative analyses in finance. Many studies applied statistical models to the stock

market.

Bachelier[6] proposed the hypothesis that stock market returns followed a normal distribution in 1900. However, subsequent research has discovered numerous instances that returns in the stock market show some characteristics of leptokurtosis, or ”fat tails”, making the fit with a normal distribution, according to Bachelier’s hypothesis, less than ideal. This also renders statistical inferences based on the normal distribution ineffective. Further research indicates that returns can be considered to follow a stable distribution[7, 8]. Benoît B. Mandelbrot studied the price of cotton, and he discovered the price of cotton could fit in the model of stable distribution, a class of heavy-tailed distributions first proposed by French mathematician Paul Lévy[9]. Its parameter determines the shape of the distribution, with $\alpha = 1.7$ indicating that the cotton price returns exhibit high peaks and heavy tails. Compared to the normal distribution, cotton prices are more likely to have extreme changes. Still, the many parameters in a stable distribution make the stable distribution model difficult to apply in analysing the stock market and its returns[10].

Other than fitting returns into a certain distribution, Survival analysis(SA)[11], a fundamental family of statistical models primarily focusing on predicting the time until the occurrence of an event, can also be used to analyse the stock market. Originating from biomedical sciences, where it was employed to analyze the time until death or failure, survival analysis has found applications across a myriad of domains due to its ability to handle ’time-to-event’ data while accounting for censoring and other complex issues. Within a financial context, the principles of survival analysis adopt a slightly different, yet equally significant, guise. Instead of analyzing biological survival, researchers shift their attention to the survival of assets. In this realm, the ”event” refers to various financial occurrences that can impact an asset’s ’life’ or ’health’. Examples might include a bond defaulting, a stock reaching a certain unfavourable threshold, or a bank facing liquidity crunches[12].

In line with Fama’s Efficient Market Hypothesis[13], it is posited that market actors swiftly apprehend all pertinent information and promptly incorporate it into market valuations. The current prices in the capital markets fully reflect all publicly available information. Also, as Fama proposed that today’s returns are independent and unrelated to yesterday’s returns, implying that price changes follow a random walk process[14]. Any individual who neither has better insights of the insight of the market should randomly price the asset rather than use any fundamental analysis tools. However, some studies have refuted the random walk theory. One line of research suggests that stock

markets exhibit a positive correlation in the short term and a negative correlation in the long term[15].

While asset returns have been extensively analyzed, less attention has historically been paid to modelling the survival dynamics of equities themselves. Standard risk models often assume stocks have indefinite lifespans, leading to underestimating the risk-related events in the stock market, such as delisting risk[16].

Survival analysis stands as a pivotal branch of statistics, especially when delving into the probability of occurrence of an event over time. At its core, survival analysis endeavours to predict and analyze the time until an event of interest occurs, which, in financial contexts, might refer to events such as loan defaults, stock market crashes, or bankruptcy filings. However, unpacking the different types of survival analysis reveals a terrain marked with intricate technical complexities[17].

When we probe deeper into the analysis of survival-related events, we encounter an intricate landscape punctuated with challenges. These technical intricacies manifest in various forms, such as censoring, competing risks, and time-varying covariates[17]. Each adds a layer of complication to the modelling process, often making it a formidable task to produce reliable and actionable insights.

Recognizing these challenges, this project's primary aspiration is to develop a robust empirical platform tailored specifically for this realm. This platform is envisioned to not only address the inherent complexities of survival analysis but also to streamline the process of experimentation. The idea is to facilitate exhaustive analyses of financial-related risks within the survival analysis framework, allowing for a more nuanced understanding and prediction of risk events over time.

In short, some key motivations in this project for applying survival analysis to stock include:

1. **Significance in Stock Market Analysis:** Survival analysis offers valuable insights into stock market behaviors, such as understanding how long a particular stock might sustain a bullish trend or predicting the longevity of bearish phases.

2. **Incorporating Financial Indicators:** Stocks are influenced by a plethora of financial indicators. Our platform would seamlessly integrate these, allowing for a nuanced analysis of their impact over time.

3. **Democratizing Complex Analysis:** While survival analysis has profound applications, its complexity can deter many from utilizing it. The platform aims to

simplify this process, making it accessible to even those without a deep statistical background.

4. **Interactive Experimentation:** An intuitive interface would allow users to experiment, tweak parameters, and visualize outcomes, promoting a more hands-on approach to the stock market analysis.

5. **Enhancing Predictive Modelling:** The platform would offer a multi-faceted approach to stock market forecasting by integrating survival analysis results with other predictive tools.

Based on the above, Survival analysis can provide a valuable framework for risk managers, quantitative investors, and financial economists seeking better to understand corporate longevity and dissolution dynamics and drivers. It allows richer, more realistic modelling of the lifecycle of listed equities.

1.2 Project Objective

As addressed in the section1.1, the primary objective is to develop a web interface that will combine survival analysis with the dynamic nature of the stock market.

Survival analysis stands apart from many conventional statistical models in its data representation and preprocessing requirements. While standard models typically handle cross-sectional or time-series data, survival analysis delves into 'time-to-event' data. This type of data captures not only the length of time leading up to an event but also indicates if the event was observed (uncensored) or left unobserved (censored) by the study's conclusion.

Such data specificity demands meticulous preprocessing. Key among these steps is managing censored observations, which arise when the event of interest either remains unobserved at the study's end or if participants exit the study prematurely. Challenges like accounting for time-varying covariates, navigating competing risks, and addressing other complexities complicate data preparation and modelling stages.

These nuanced demands create a considerable barrier in terms of skillset. To adeptly navigate survival analysis, practitioners must deeply grasp its theoretical foundation as well as the data's intricacies. This creates an economic challenge for organizations lacking this expertise in-house, as they face potential costs in recruiting or upskilling data professionals. Consequently, even with the profound insights survival analysis might yield, its intricate demands can hinder its broader application in some industries.

However, our web interface is crafted to simplify this process. Designed for clarity and ease of use, it aims to streamline data collection and model execution. This design choice ensures survival analysis becomes more accessible, even for investors without a statistical foundation.

More specifically, this project will have several key objectives:

1. **Data Collection and Processing**: This project will implement a robust data collection mechanism to fetch relevant financial data from various sources. The application will be designed to handle tasks such as data cleaning, transformation, and structuring, which are crucial for conducting survival analysis. The covariates to be considered for the analysis will be chosen based on a thorough literature review, industry practices, and the feasibility of data collection.

2. **Survival Analysis Implementation**: This project will develop a survival analysis model that is well-suited to incorporate the survival time of financial market risks. The model should be capable of handling complexities such as censored data and should take into account factors like market conditions, stock volatility, and financial indicators. Both parametric and non-parametric methods will be explored.

3. **Web Interface Design and Development**: A primary focus of the project is the development of an intuitive and user-friendly web interface. The design will allow users to easily select their stocks of interest, adjust parameters as needed, and conduct survival analysis with a simple click.

4. **Visualization and Interpretation**: The web interface will be equipped with visualization tools to present the results of the survival analysis clearly. This will include survival curves, hazard functions, and confidence intervals. Furthermore, the interface will provide interpretations of the results to aid users in making informed investment decisions.

5. **Usability Testing and Validation**: Rigorous usability testing and validation will be conducted to ensure that the web interface operates as expected and is user-friendly. Feedback will be solicited from test users to refine the interface and improve its functionality.

Through these objectives, this project seeks to break down the barriers to the application of survival analysis in stock market investment and provide a tool that

investors can use to improve their decision-making process. As a secondary goal, this project also aims further to validate the potential of survival analysis in financial markets, contributing to the body of knowledge in this area.

1.3 Novelty and Significance

Several novel features characterize this project and hold considerable significance in the domain of finance and technology. Here are the key elements:

1. **Integration of Survival Analysis and Stock Market**: Despite the proven potential of survival analysis in various fields, its application in the realm of the stock market is relatively less explored. This project introduces a new way of predicting and understanding stock market behaviors using survival analysis, bringing in a fresh perspective to investment strategies.

2. **User-friendly Interface**: The development of a user-friendly web interface to perform survival analysis on the stock market is a novel aspect of this project. It aims to bridge the gap between advanced statistical techniques and users with no or limited statistical knowledge.

3. **Automated Data Collection and Processing**: The automation of data collection and preprocessing, which is a necessary and often cumbersome step in any analysis, is an innovative feature. By incorporating this, the project simplifies the process, making survival analysis more accessible to a broad audience.

4. **Comprehensive Visualization Tools**: The inclusion of comprehensive visualization tools for presenting the results of survival analysis is a distinctive feature. It ensures that the analysis outcome can be easily understood by users, regardless of their level of statistical proficiency.

5. **Market Significance**: The interface's capability to guide investors in making informed decisions by providing insights into the time-dependent behavior of stocks is of high market significance. It could potentially impact how individual and institutional investors approach their investment strategies.

1.4 Dissertation Structure

The remainder of this paper is organised as follows. Chapter 2 presents the literature relating to different types of SA models and the application of SA models in analysing stock market behaviours. Chapter 3 details the methodology used in this research, discussing the data sets leveraged, the preprocessing steps undertaken, and the model setups, including the Parametric Regression Model, Cox proportional hazard model, and Non-Parametric Models. Chapter 4 delves into the visualization and its applications, elucidating the web interface's project environment and design aspects, covering areas such as the welcome page, delisting analysis, and the value at risk analysis.

Chapter 2

Related work

2.1 Survival analysis

Survival analysis(SA)[18] is a branch of statistics methods that analyses the expected duration of time until one or more events happen. Generally, survival analysis aims to estimate survival and hazard functions from data. The survival function, $S(t)$, gives the probability that an individual survives longer than some specified time t. The hazard function, $h(t)$, gives the instantaneous potential per unit time for the event, given that the individual has survived up to time t. Generally, there are three types of SA: non-parametric, semi-parametric, and parametric.

When fitting actual survival data to distributions, these models can be fitted separately, and the model with the best fit is selected based on goodness-of-fit tests. Sometimes, for a set of survival data, the exact distribution of the survival time is unknown in advance, and it is also difficult to determine which distribution is most suitable. In this case, nonparametric or semi-parametric regression models can be used. However, if it is known that a set of data does follow a certain parametric distribution, it is better to use the corresponding parametric regression model, as parametric methods generally have higher accuracy than nonparametric or semi-parametric methods [**?**].

An important aspect of parametric regression model analysis for survival data is model fitting or distribution fitting. Common probability distributions to describe survival data include exponential, Weibull, log-logistic, lognormal, and generalised gamma distribution models. These distributions use probability density function $f(t)$, survival function $S(t)$, and hazard function $h(t)$ (or hazard rate function) to describe survival data [18]. If one of the functions is given, the other two functions can be derived, and their relationships are:

1. If $f(t)$ is known, then $S(t) = \int_t^\infty f(u)du$, $h(t) = \frac{f(t)}{S(t)}$
2. If $S(t)$ is known, then $f(t) = -\frac{d[S(t)]}{dt}$, $h(t) = \frac{f(t)}{S(t)} = -\frac{d[\ln S(t)]}{dt}$
3. If $h(t)$ is known, then $S(t) = \exp(-\int_0^t h(u)du)$, $f(t) = h(t)S(t)$

Various models are proposed within the SA family, such as the Kaplan-Meier estimator[19], log-rank test[20], Cox proportional hazards regression(Cox PH) model[21], and parametric models[22]. Each has advantages and disadvantages depending on the goals and characteristics of the data set. SA can account for incomplete observations and censoring to provide valid statistical inferences. Non-parametric models such as Kaplan-Meier estimator do not make any assumption to the distribution of survival time, making it considered the universal model for survival analysis. Semi-parametric models, such as the Cox PH model, is based on the proportional hazards assumption. This is a fundamental premise for the Cox Proportional Hazards model. It implies that the hazard rates of different groups (or, for continuous variables, the hazard rate at one value of the variable compared to the hazard rate at a reference value) are proportional over time. The parametric models can provide more accurate prediction to the survival probability only when the survival time fits its types of distribution[22].

Modified SA has been applied in various ways to study the stock market dynamics. Constantin and Das Sarma [23] used survival analysis to study the temporal fluctuations in time-dependent stock prices. They analyzed stock price fluctuations as a non-Markovian stochastic process using the first-passage statistical concepts of persistence and survival. They found that the fluctuating stock price is multifractal, and the choice of the sampling time has no effect on the qualitative multifractal behaviour displayed by the generalized Hurst exponent associated with the power-law evolution of the correlation function. Sandoval Junior [24] used survival analysis to study the cluster formation and evolution of 92 indices of stock exchanges worldwide from 2007 to 2010, which includes the Subprime Mortgage Crisis of 2008. The study focused on the survivability of connections and of clusters through time and the influence of noise in centrality measures applied to the networks of financial indices. Wergen [25] studied the statistics of record-breaking events in daily stock prices of 366 stocks from the Standard and Poors 500 stock index. The study found that the number of records in the stocks appears to be systematically decreased compared to the random walk model. Scalas et al. [26] applied continuous-time random walks (CTRWs) as phenomenological models of the high-frequency price dynamics. Their empirical analysis of the 30 DJIA stocks shows

that the waiting-time survival probability for high-frequency data is non-exponential. This fact sets limits for agent-based models of financial markets.

As discussed in this section, the family of SA models has been used to study temporal fluctuations, cluster formation and evolution, record-breaking events, and high-frequency price dynamics. This shows how that survival analysis can be deployed in various ways to study the stock market dynamics while providing valuable insights into the behaviour of the stock market and can inform investment strategies and decision-making processes.

2.2 Technical Analysis in Stock Market and Value at Risk

Early work by Fama and Blume in the 1960s found support for the random walk hypothesis and weak evidence for technical analysis[14] as we mentioned in section1.1. However, Brock et al. in 1992 demonstrated the success of moving average and trading breakout rules on the Dow Jones Industrial Average, challenging the Efficient Market Hypothesis (EMH)[27]. Additional backtesting found momentum indicators like the relative strength index (RSI) produced significant excess returns[28].

In the 2000s, the adaptive market hypothesis emerged as an alternative to the EMH. Lo et al. argue markets evolve dynamically, and inefficiencies exist temporarily as people adapt beliefs[29]. Evidence for the profitability of technical trading continued to accumulate in international markets[30, 31]. With the rise of algorithmic and high-frequency trading (HFT), technical strategies leveraging machine learning have become more sophisticated. Kara et al. combined multiple technical indicators with neural networks to forecast price direction[32]. Krauss et al. optimized technical rules with genetic algorithms, finding significant out performance[33].

Value at Risk (VaR) has become a cornerstone in the realm of financial risk management since its inception. Originating in the late 1980s and early 1990s, the need for a more systematic approach to risk management was underscored by significant financial downturns, notably the stock market crash of 1987[34]. This event highlighted the vulnerabilities in the financial sector, prompting a search for more robust risk measurement tools. VaR emerged as a solution, offering a quantitative estimate of the potential loss an investment portfolio might face over a specified period for a given confidence interval.

The methodologies for calculating VaR are diverse, each with its unique approach and assumptions[35]. The historical simulation, for instance, relies on re-sampling historical returns to generate a distribution of potential future returns, operating on the

premise that past patterns will recur. On the other hand, the variance-covariance method is rooted in the assumption that returns are normally distributed, utilizing the mean and standard deviation of returns to estimate VaR[35]. The Monte Carlo simulation, arguably the most intricate of the three, generates a multitude of random price paths for assets in a portfolio to predict potential future returns[35].

In the stock market, VaR's influence is undeniable. Financial institutions have integrated VaR into their operations to determine capital reserves, set trading limits, and even allocate capital among various trading desks[36]. Its significance is further emphasized by regulatory bodies that have woven VaR calculations into their capital adequacy requirements for banks and other financial entities.

2.3 Fundamental Analysis in Stock Market

Fundamental analysis stands as a cornerstone in the evaluation of a company's intrinsic value, leveraging financial statements such as balance sheets and income statements to make informed investment decisions. This section delves into the methodologies and insights derived from these financial documents, emphasizing their significance in the stock market.

2.3.1 Fundamentals

In the intricate world of stock market analysis, the balance sheet, income statement, and fundamentals serve as foundational pillars, offering a comprehensive view of a company's financial health and performance[37]. When used adeptly, these tools can guide investors in making informed decisions, assessing a company's value, and predicting its future trajectory.

The balance sheet provides a snapshot of a company's financial position at a specific moment in time, detailing its assets, liabilities, and shareholders' equity[38]. It's akin to a financial photograph, capturing the company's resources, obligations, and the residual interest left for shareholders after all debts are settled. Investors can gauge a company's financial stability, liquidity, and risk profile by examining a balance sheet. For instance, a high debt-to-equity ratio might indicate a company's heavy reliance on debt, potentially signalling increased financial risk.

On the other hand, the income statement offers a dynamic view of a company's financial activities over a period, be it a quarter or a year[39]. It chronicles the journey

of a company's earnings, charting its revenues, expenses, and the resulting net profit or loss. This statement is pivotal for understanding a company's profitability. A consistent track record of increasing revenues and controlled expenses can be a positive sign, indicating a company's growth trajectory and efficient operations.

In the realm of stock market analysis, these tools collectively offer a holistic perspective. They allow analysts and investors to assess a company's financial health, evaluate its profitability, and make informed investment decisions. Moreover, they serve as a compass, guiding stakeholders in forecasting a company's future performance based on historical data, industry trends, and economic indicators.

2.3.2 Altman Z Score

The Altman Z-score, introduced by Dr Edward I. Altman in 1968[40], serves as a seminal financial metric meticulously crafted to prognosticate the likelihood of a firm's insolvency within a two-year horizon. The Altman Z-score's robustness and adaptability have been the subject of extensive scholarly attention, with researchers applying the model across a diverse array of sectors and geographies.

One of the seminal studies in this domain focused on the coal mining companies in Indonesia during the period 2012-2016[41]. This research, which employed the Altman Z-Score methodology, found that several entities were categorized into potential bankruptcy, grey zones, and financially robust segments[41]. This study underscores the Z-score's versatility in its applicability across diverse sectors, even those with unique financial dynamics like the mining industry.

Further emphasizing the model's adaptability, a comprehensive financial assessment of Maruti Suzuki Ltd was conducted over a decade, leveraging the Altman Z-score model[42]. The findings from this study provided a nuanced understanding of the fiscal health of specific corporate entities, highlighting the model's diagnostic capabilities.

In a more global context, the Altman Z-Score's resonance was evident in a study that scrutinized the financial distress status of select NIFTY 50 entities in the Indian Stock Market[43]. Such research accentuates the model's universal applicability, transcending regional financial idiosyncrasies.

Over the years, the Altman Z-score model has undergone iterative refinements. Its applications have evolved beyond mere bankruptcy prediction, as evidenced by a study exploring external analytical paradigms and introspective evaluations of financially beleaguered firms[44]. This research provided a holistic view of the model's utility,

spanning diverse analytical frameworks.

2.4 Macroeconomics

The application of macroeconomic indicators in stock market analysis has been the subject of extensive research. These indicators provide valuable insights into the overall economic conditions that can influence stock market performance.

A study by Khan and Billah[45] focused on the Dhaka Stock Exchange, examining the relationship between the stock exchange index return and macroeconomic variables such as exchange rate, inflation, and money supply. They used the Johnson Cointegration test, Augmented Dicky Fuller (ADF), and Phillip Perron (PP) tests to analyze the long-term relationship between these variables and stock market returns. Their findings revealed a strong association between stock returns and the consumer price index, money supply, and exchange rates. Interestingly, they found a negative association between market capitalization and stock returns, suggesting that the Dhaka stock exchange is reactive to macroeconomic indicators in the long run.

In a different context, Bock [46]investigated the potential of using Google query volumes for related search terms to predict changes in the U.S. unemployment rate before the official news release. He found that this approach improved the predictability of the U.S. unemployment rate and enhanced market timing of trading strategies when considered jointly with macroeconomic data. This study illustrates the potential of combining extensive behavioural data sets with economic data to anticipate investor expectations and stock market moves.

In essence, these studies collectively highlight the profound influence of macroeconomic indicators on stock market behaviour and underscore the potential of innovative data-driven approaches in enhancing market predictions.

Chapter 3

Data & Methodology

3.1 Data-sets

This study requires an extensive and diverse dataset of stock information to facilitate the analysis and modelling of diverse financial assets. Historical stock market data will be procured via the Yfinance API[47], a widely adopted and robust tool for accessing historical market data from Yahoo Finance, chosen for its comprehensive data offerings, user-friendly interface, and established reliability. The Yfinance API provides comprehensive historical data for publicly traded companies, encompassing daily metrics such as opening, closing, maximum, and minimum stock prices and daily trading volume. Stock historical data will be extracted with the maximum available temporal span, the table3.1 shows the snapshot of the stock historical of **AAPL**. The `Date` column lists the specific date of the stock trading activity using the format of YYYY-MM-DD. The `Open`column shows the opening price of the stock on the corresponding date. `High`indicates the highest price the stock reached during the trading session of the corresponding date. Similarly, `Low` represents the lowest price the stock reached during the trading session of the mentioned date. `Close` provides the stock's closing price on the given date. It's the last price at which the stock was traded during that particular trading session. `Volume` denotes the number of shares that were traded during the trading session of the corresponding date. The values are given in scientific notation. Finally, `Dividends` shows if any dividends were distributed on a particular date.

Additionally, Yfinance offers datasets containing companies' fundamental information, including income statements, balance sheets, and common accounting ratios. Balance Sheet Data offers a snapshot of a company's financial health at a specific point

Date	Open	High	Low	Close	Volume	Dividends
1980-12-12	0.128348	0.128906	0.128348	0.128348	4.76×10^9	0.0
1980-12-15	0.122210	0.122210	0.121652	0.121652	1.88×10^9	0.0
1980-12-16	0.113281	0.113281	0.112723	0.112723	1.024×10^9	0.0
1980-12-17	0.115513	0.116071	0.115513	0.115513	5.20×10^8	0.0
1980-12-18	0.118862	0.119420	0.118862	0.118862	3.248×10^8	0.0

Table 3.1: Stock Historical Price Data set

in time, revealing metrics such as total assets, liabilities, shareholder equity, and current ratios, providing insights into financial stability and short- and long-term obligations. Income Statement Data reflects a company's financial performance over a defined period, analyzing revenues, expenses, and profits to discern profitability trends and overall sustainability. Fundamentals encompass a comprehensive range of data points, including market capitalization, earnings per share (EPS), price-to-earnings (P/E) ratios, and dividend yields, offering a holistic valuation perspective. As this study does not focus on time-varying variables' impact on SA models and risk analysis, only the most recent Balance Sheet, Income Statement, and Fundamentals data for the specified companies will be collected. Consequently, the Balance Sheet, Income Statement, and fundamental data will be collected from Yfinance to underpin the estimation of pivotal SA models within this study.

This project will also use AlphaVantage[48] to acquire the company's listing and delisting status information. AlphaVantage is a platform that offers a comprehensive suite of stock market data APIs. It provides developers and investors with a set of tools to fetch real-time and historical stock market data, forex (foreign exchange) rates, cryptocurrency data, and other financial metrics. AlphaVantage listing and delisting API will be used in this project to acquire a comprehensive list of US-listed equities, both active and delisted, covering the full history of US markets. Researchers can retrieve point-in-time snapshots to conduct analysis on asset survival rates, corporate events, index reconstitutions, and other areas of interest. By covering the complete lifespan of securities, this API enables robust backtesting and insights into market dynamics over time. Whether analyzing index composition changes, modeling survivorship bias, or tracking equity lifecycles, the Historical Stocks and ETFs API provides the core dataset for research on US markets over time. This project will use the delisting and active

Symbol	Exchange	Asset Type	IPO Date	Delisting Date
AA-W	NYSE	Stock	2016-10-18	2016-11-08
AAAP	NASDAQ	Stock	2015-11-11	2018-02-20
AABA	NASDAQ	Stock	1996-04-12	2019-11-06
AAC	NYSE	Stock	2014-10-02	2021-04-19
AACC	NASDAQ	Stock	2004-02-05	2013-10-17

Table 3.2: Listing and Delisting Data Set

stock data to calculate the survival time for every delisting company for SA model. The data set returned form AlphaVantage is shown in table 3.2, where `Symbol` column displays the unique ticker symbol associated with each stock. `Exchange` indicates the stock exchange where the stock is (or was) listed. This project collects data from NYSE (New York Stock Exchange), NYSE MKT, NYSE ARCA and NASDAQ. `Asset Type` specifies the type of financial asset. This project will filter out any asset that is not an asset type of stock. `IPO Date` IPO stands for Initial Public Offering. This column provides the date on which the stock was first made available to the public for purchase. It's the day the company went public. `Delisting Date` This is the date on which the stock was removed or delisted from its respective stock exchange. A stock can be delisted for various reasons, including failing to meet the exchange's criteria, bankruptcy, mergers, or acquisitions.

3.2 Prepocessing

3.2.1 Delisting risk analysis data

This project will define the 'Survival Event' as stock delisting, and the duration time is the time difference between the initial public offer and the delisting date, which can be calculated using the formula3.3, where $I = 0$, if stock is currently active and $I = 1$, is stock is delisted.

$$\mathbf{I} = \begin{bmatrix} I \\ 1 - I \end{bmatrix} \tag{3.1}$$

$$\mathbf{T} = \begin{bmatrix} T_{\text{delisting}} \\ T_{\text{current}} \end{bmatrix} \tag{3.2}$$

$$T_{survival} = \mathbf{I}^{\top}\mathbf{T} - T_{IPO} \tag{3.3}$$

According to Dong's research[49] for using SA models to study the stock delisting in China's stock market, he proposed that some time-varying covariates(TVCs) can be used as covariants in Cox proportional hazard model. As shown in the table**??**. Some of the TVCs will not be able to acquire from yfiancne directly. This project calculates these TVCs according to their definition.

Current Asset Turnover Ratio (CATR) measures how efficiently a company utilizes its current assets to generate sales revenue. It is calculated by dividing net sales by the average current assets and defined as follows:

$$\text{CATR} = \frac{\text{Net Sales}}{\text{Average Current Assets}} \tag{3.4}$$

Total Asset Turnover Ratio (TATR) evaluates a company's efficiency in utilizing all its assets to produce sales. It is computed by dividing net sales by average total assets:

$$\text{TATR} = \frac{\text{Net Sales}}{\text{Average Total Assets}} \tag{3.5}$$

The current Ratio (CR) gauges a company's capability to pay off its short-term obligations with its short-term resources. It is derived by dividing current assets by current liabilities. The formula is:

$$\text{CR} = \frac{\text{Current Assets}}{\text{Current Liabilities}} \tag{3.6}$$

Net Profit Ratio (NPR) represents the percentage of net income generated from net sales. It is calculated by dividing net income by net sales:

$$\text{NPR} = \frac{\text{Net Income}}{\text{Net Sales}} \tag{3.7}$$

Earnings Per Share (EPS) indicates the portion of a company's profit allocated to each outstanding common stock share. It is computed by dividing net income by the number of outstanding shares. The formula is:

$$\text{EPS} = \frac{\text{Net Income}}{\text{Number of Outstanding Shares}} \tag{3.8}$$

The Altman Z-Score predicts the likelihood of a company going bankrupt as we introduced in the section2.3.2. It considers five financial ratios weighted differently:

$$AltmanZ - Score = 1.2A + 1.4B + 3.3C + 0.6D + 1.0E \tag{3.9}$$

Where:

- A = Working Capital / Total Assets
- B = Retained Earnings / Total Assets
- C = Earnings Before Interest and Taxes / Total Assets
- D = Market Value of Equity / Total Liabilities
- E = Sales / Total Assets

This project will create a database in which each row records a company's survival time and all TVCs calculated. This database will later be used to analyse the risk of company delisting. Table 3.3 shows the statistical information for all TVCs calculated according to equations 3.4 to 3.9.

Variable Name	**Mean**	**Max**	**Min**	**Std**
CATR	2.239	79.69	-3.728	3.013
FATR	4.692	1.087e+04	-957	169.7
TATR	0.5962	29.46	-4.501	0.9799
CR	7.93	2.52e+04	0.000253	313.6
QR	7.715	2.52e+04	-4.189	317
NPR	-11.83	1246	-4906	161.5
EPS	13.3	2.821e+04	-3.957e+04	794.4
DE	2.869	3614	-2131	82.72
Altman Z-Score	1.931	292.5	-459.4	14.09

Table 3.3: Financial Metrics for active and delisting companies in records

3.2.2 Value at Risk Analysis

To apply the SA to VaR analysis result, the data processing aims to pinpoint the first instance when a selected stock price breaches a calculated Value-at-Risk (VaR) over a rolling window of 3 months as many studies mentioned in section 2.2 have chosen. To simplify the data process, this section will explain the process with a confidence level of 95%, which is commonly used in VaR analysis and has also been used in many studies introduced in the section2.2. The pseudocode is also attached in the appendixA.1.

By defining two key terms, $P(t)$ denotes the stock's closing price on day t. The second, $VaR(t)$, represents the Value-at-Risk — a measure indicating the potential loss

over a set period for a given confidence interval. Consider a one-day VaR estimation of $1 million at a 95% confidence level to elucidate. This interpretation suggests a 5% likelihood that the portfolio will experience a decline in value exceeding $1 million within a single day, assuming typical market conditions.

Mathematically, for a given confidence level α (represented as 95% in this project), and over a defined time horizon t, $VaR(t)$ is given by:

$$VaR_{\alpha}(t) = \inf\{l : P(L > l) \leq 1 - \alpha\} \tag{3.10}$$

In this equation:

- L denotes the potential portfolio loss over the time horizon t.
- $P(L > l)$ expresses the probability that the loss will surpass the value l.

This project will propose a methodology to estimate the potential loss in stock prices over a given period. The closing price of a stock at time t is represented by $P(t)$

Using historical stock price data from the three months (approximately 63 trading days)preceding t, a Monte Carlo simulation is executed to forecast prices for the subsequent three months, resulting in the sequence:

$$\{P(t+1), P(t+2), \ldots, P(t+63)\} \tag{3.11}$$

The 95% Value at Risk (VaR) at time t is then computed as:

$$\text{VaR}(t) = P(t) - \text{percentile}_{5\%}(\{P(t+1), P(t+2), \ldots, P(t+63)\}) \tag{3.12}$$

Or in other words, Every day that ends our rolling 3-month window becomes a t. For each such t, we exam into the past 3 months of stock returns and play a Monte Carlo simulation which will be introduced in the section3.2.2.1, simulating possible stock prices for the upcoming 3 months. These simulations produce a distribution of stock returns. From the simulated distribution of stock returns, we are particularly interested in the lowest 5th percentile of prices p, termed as $P_{5\%}$. The gap between the actual stock price on day t and this 5% threshold gives us our $VaR(t)$.

To identify potential risk, each day d from $t+1$ to $t+63$ is examined. If the following condition is met:

$$P_{5\%} < P(t) - \text{VaR}(t) \tag{3.13}$$

then d is considered a VaR breach event. The risk exposure duration for each breach event is calculated as follows:

$$\text{duration} = d - (t+1) \tag{3.14}$$

3.2.2.1 The Monte Carlo Simulation

To simulate the price of a selected stock, this project will utilize the Geometric Brownian Motion (GBM). It is a commonly employed stochastic process for describing the evolution of stock prices.

The discrete form of the GBM equation is given by:

$$P_{t+1} = P_t \times (1 + \mu\Delta t + \sigma\varepsilon\sqrt{\Delta t}) \tag{3.15}$$

Where:

- P_t is the stock price at time t.
- μ is the expected daily return.
- σ is the standard deviation of stock returns (volatility).
- ε is a random number drawn from a standard normal distribution.
- Δt is the time step, typically 1 day.

Repeat the above stochastic process numerous times (for instance, 10,000 times) to simulate various trajectories of stock prices. Each trajectory represents a potential future evolution of the stock price.

3.3 Models Setup

3.3.1 Parametric Regression Model

The figure3.1 shows the distribution of survival time of active and delisting companies used in this project, respectively, and figure **??** shows the distribution combining both delisting and active companies. As mentioned in section 2.1, parametric regression models can provide better results if we can identify a certain type of survival time distribution.

It is not hard to identify the survival time distribution, as visualized in the histograms with KDE in figure3.1, appears to be right-skewed. The majority of companies have a shorter survival time, with fewer companies surviving for extended periods. However, to apply SA parametric regression model, this project will first analyse the type of survival time distribution.

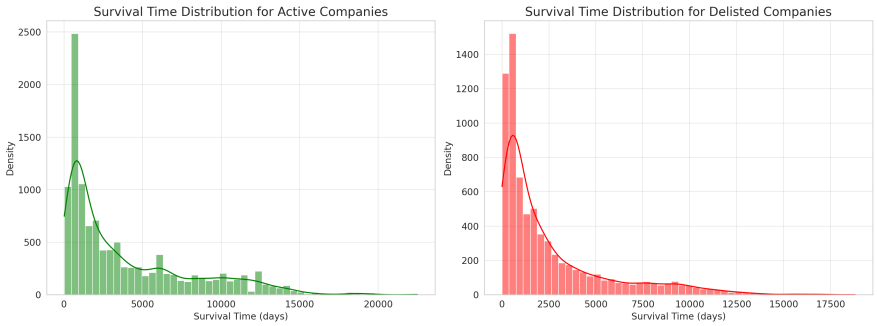


Figure 3.1: Survival Time Distribution for Active and Delisted Companies

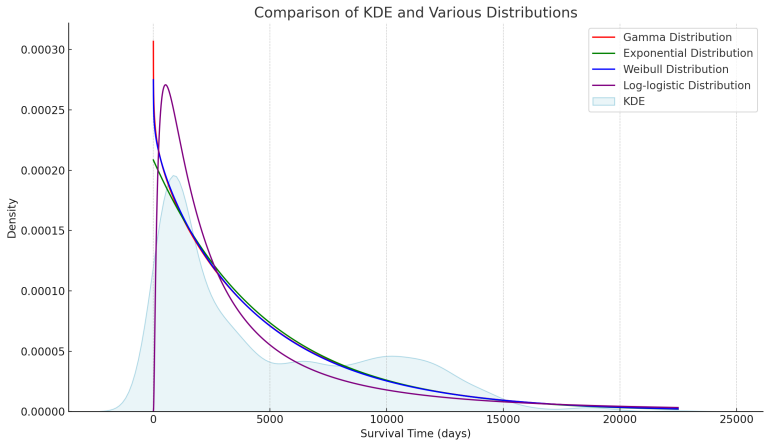


Figure 3.2: Comparison of KDE and Various Distribution

The figure3.2 compares the KDE of the survival time with other types of distribution.

The exponential, gamma, Weibull, and log-logistic distributions are fundamental continuous probability distributions frequently employed in stock market analysis. Each distribution is characterized by specific parameters that shape its form and characteristics. By fitting a distribution to data and estimating these parameters, we can gain insights into the underlying patterns and tendencies of the data and also will be able to select the best survival time distribution as needed for the parametric SA model.

Mathematically, The PDF of the Gamma distribution is given by:

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \tag{3.16}$$

where:

- α is the shape parameter.
- β is the scale parameter.
- Γ denotes the Gamma function.

Based on the fitting of our active and delisted companies' survival time, the estimated parameters are:

- α (Shape): 0.944
- β (Scale): 5082.79

The PDF of the Exponential distribution is defined as:

$$f(x;\lambda) = \lambda e^{-\lambda x} \tag{3.17}$$

where λ is the rate parameter. This parameter is the inverse of the scale parameter: $\lambda = \frac{1}{\text{scale}}$. Based on the fitting, the estimated parameter is:

- λ (Rate): 0.00020846

The PDF of the Weibull distribution is described by:

$$f(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \tag{3.18}$$

where:

- k is the shape parameter.
- λ is the scale parameter.

Based on the fitting, the estimated parameters are:

- k (Shape): 0.965
- λ (Scale): 4722.38

The PDF of the Log-logistic distribution is:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \tag{3.19}$$

where:

- μ represents the mean of the natural logarithm of the variable.
- σ is the standard deviation of the natural logarithm of the variable.

Based on the fitting, the estimated parameters are:

- μ (Shape): 1.264

- σ (Scale): 2591.39

This detailed analysis provides us with the necessary parameters to model the survival time of companies using the aforementioned distributions.

Upon fitting these distributions to the survival time data, visual inspection suggested that the Gamma and Weibull distributions provided a good fit. This project will also use the K-S test to compare a sample's empirical distribution function (ECDF) with the cumulative distribution function (CDF) of a specified theoretical distribution. The K-S test statistic is:

$$D = \max_x |F_n(x) - F(x)| \tag{3.20}$$

Where $F_n(x)$ is the ECDF and $F(x)$ is the CDF of the specified distribution.

K-S test can be taken as a structured approach to discern the most appropriate distribution for data wherein the "survival time" exceeds 300 days. The K-S statistic encapsulates the maximal discrepancy between the two. A small K-S value signifies a more commendable fit.

The acquired results show that the Gamma distribution offers the most favourable fit to the data, given its minimal K-S statistic. The Exponential distribution closely follows this, while the K-S statistics for the Weibull and Log-logistic distributions are marginally higher as all D is shown in equation 3.21 to 3.24.

$$D_{\text{Exponential}} = 0.0972 \tag{3.21}$$

$$D_{\text{Gamma}} = 0.0868 \tag{3.22}$$

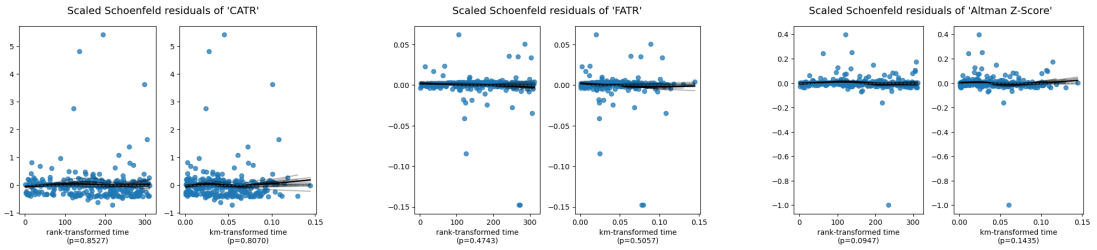
$$D_{\text{Weibull}} = 0.1125 \tag{3.23}$$

$$D_{\text{Log-logistic}} = 0.1084 \tag{3.24}$$

Thus, for the survival time in our project, the Gamma and Exponential distributions appear to provide the optimal fit and will be implemented into the web interface.

3.3.2 Cox proportional hazard(Cox PH) model

Cox PH is a semi-parametric model, as introduced in the section2.1. This project will use fundamentals calculated from the company's balance sheet and income statement as covariates, and the survival time is the time difference between IPO and time of delisting as stated in the equation 3.3. Lifelines provide a function to directly fit the Cox PH model by passing the Panda DataFrame, event, and duration as parameters.



(a) CATR (b) FATR (c) Altmen Z-Score

Figure 3.3: Test Results of PH Assumption

The PH assumption, introduced in the section2.1, will also be tested by inspecting the Schoenfeld residuals as suggested by the lifelines documentation[50]. However, some suggest that checking the PH assumption is not necessarily a precondition to applying the Cox PH model when researchers do not care about the propositional hazards and the only goal is survival prediction[51]. To show how the scaled Schoenfeld residuals work when checking the assumption, this project also conducts a test of PH assumption and the results are shown in the figure3.3.

3.3.3 Non Parametric Models

Traditional survival analysis models, such as the Cox proportional hazards model, make specific assumptions about the underlying hazard function. In contrast, non-parametric models of survival analysis make no such assumptions about the form of the hazard function, which makes them more flexible and can handle non-proportional hazards. Additionally, deploying a non-parametric model can provide a robust result for the Breach of VaR analysis, considering that the survival time distribution can vary from different date windows. This project will deploy the Kaplan-Meier estimator and Nelson-Aalen estimator for the Breach of VaR analysis.

The Kaplan-Meier (KM) estimator estimates the survival function $S(t)$ without assuming a specific functional form for the survival distribution.

The Kaplan-Meier estimate of the survival function is given by:

$$S(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.25}$$

Where:

- t_i are the distinct observed event times in ascending order.
- d_i is the number of events (e.g., deaths) at time t_i.

- n_i is the number of individuals at risk just before time t_i.

Another non-parametric estimator is the Nelson-Aalen estimator, which estimates the cumulative hazard function. The cumulative hazard function $H(t)$ is related to the survival function by $S(t) = \exp(-H(t))$.

The Nelson-Aalen estimator is given by:

$$H(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i} \tag{3.26}$$

To clearly explain the data process of fitting a non-parametric model, consider the following description of a dataset with four columns.

1. `Prediction_Start_Date`: The date at which a prediction starts.
2. `VaR_Breach_Date`: The date at which a Value-at-Risk (VaR) breach occurs.
3. `Duration_Days`: The duration (in days) between the start of the prediction and the occurrence of the VaR breach.
4. `VaR_Breach`: A boolean indicating whether a VaR breach occurred (`True`) or not (`False`).

Given this structure, this project will use the `Duration_Days` column as the time-to-event data and the `VaR_Breach` column as the event indicator for survival analysis to apply the Kaplan-Meier and Nelson-Aalen estimators as follow steps:

1. **Data Preparation**:
   - Extract the time-to-event data (i.e., `Duration_Days`) and the event indicator (i.e., `VaR_Breach`).

     $$\text{Time-to-Event} = \text{Duration\_Days}$$

     $$\text{Event Indicator} = \text{VaR\_Breach}$$

2. **Kaplan-Meier Estimation**:
   - For each unique time point t in ascending order, calculate the survival probability using:

     $$S(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

     Where:

  - t_i are the distinct observed event times in ascending order.
  - d_i is the number of events (e.g., VaR breaches) at time t_i.
  - n_i is the number of individuals at risk just before time t_i.

3. **Nelson-Aalen Estimation**:

   - For each unique time point t, calculate the cumulative hazard using:

$$H(t) = \sum_{i:t_i \leq t} \frac{d_i}{n_i}$$

   Where:

   - d_i is the number of events (e.g., VaR breaches) at time t_i.
   - n_i is the number of individuals at risk just before time t_i.

Chapter 4

Visualization and Applications

4.1 Project Environment

The Web interface of this project is primarily developed using Streamlit[52]. Streamlit is an open-source Python library designed to facilitate the development of web applications specifically for data projects. Its emergence in data science and machine learning can be attributed to the increasing need to disseminate computational results in an interactive format effectively.

A noteworthy aspect of Streamlit's design is its integration capability. It is compatible with prominent data science tools and libraries like Pandas, Numpy, Matplotlib, and Plotly. This seamless integration aids in a smooth transition from data analysis to interactive application development. Furthermore, Streamlit employs a declarative syntax for incorporating widgets, eliminating the common requirement for callback functions that are typical in other frameworks.

This project utilizes several Python packages for data processing, visualization, and analysis. Pandas is used for general data manipulation and analysis. It provides data structures and tools to make working with tabular data fast, easy, and intuitive. Plotly is used to create interactive visualizations and charts to explore trends and relationships in the data.

For survival analysis, the lifelines and scikit-survival packages are used. Lifelines provides survival analysis models such as Kaplan-Meier curves, Cox regression, and Aalen additive models needed in this project. Scikit-survival implements additional survival models, including random survival forests and comparative Cox analysis.

The yfinance package downloads historical and current market data from Yahoo Finance to obtain real-time financial data. Technical indicators are then calculated using

the ta package, which implements over 80 technical analysis indicators commonly used in finance.

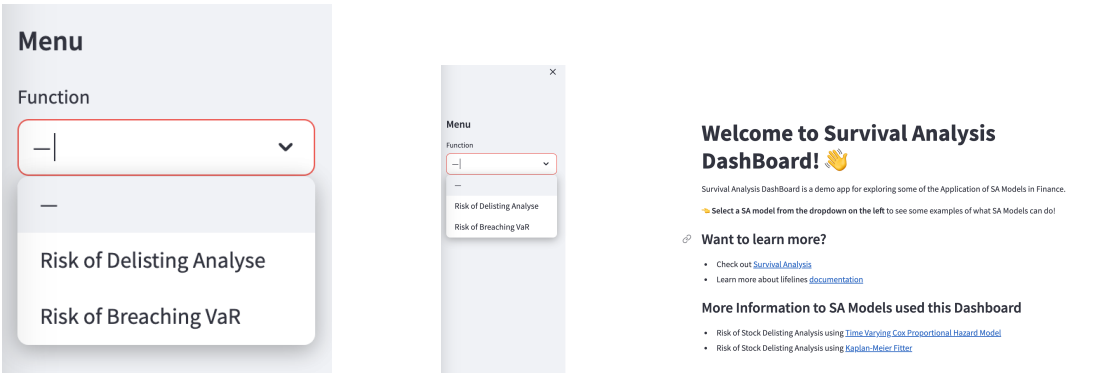
Together, these Python packages provide the tools for data wrangling, analysis, visualization, and modelling for this project. Pandas and Plotly enable exploratory data analysis while lifelines, scikit-survival, yfinance, and ta provide methods for predictive modelling and working with financial data.

4.2 Design of the Web Interface

4.2.1 The Welcome Page

In the proposed web interface, the design prominently features a sidebar on the left. This sidebar has been structured to facilitate the efficient selection of various survival analysis types. Topics such as 'delisting' and the SA for the likelihood of a 'VaR' (Value at Risk) event are among the options available for exploration. Users can simply select different types of events by clicking the drop-down menu as shown in the figure4.1a.

An organised presentation of survival analysis is provided upon transition to the main section of the welcome page, as shown in figure 4.1b. This presentation is comprehensive, aiming to introduce fundamental concepts associated with various types of survival analysis. This structured overview serves as a vital resource for general users or those unfamiliar with the nuances of survival analysis. It offers a synthesized yet detailed perspective, thereby aiding users in discerning the most appropriate survival analysis model for their respective research or projects.



(a) Menu (b) Design of the Welcome Page

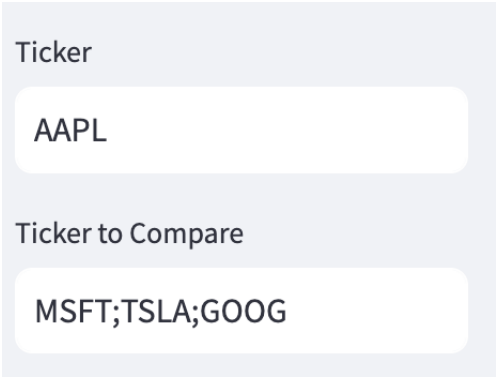
Figure 4.1: Welcome Page

4.2.2 The Delisting Analysis

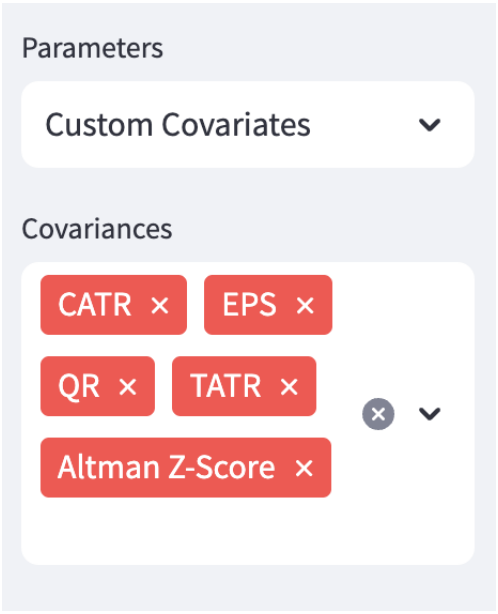
This project will utilise delisting risk analysis to demonstrate the capability of survival analysis models in handling one-time off events. The Delisting Analysis interface incorporates user-defined parameters and graphical representations to facilitate the examination of stock delistings. Users will be able to select which company to analyse and also which company to compare with as shown in figure4.2a. A drop-down menu also enables users to isolate different fundamentals, allowing a more flexible interaction with the SA model as shown in figure4.2b. To enrich the analysis, a second dropdown menu empowers users to select different models in the SA family. The table 4.1 shows the input parameters used in the delisting analysis.

Input Box Name	**Description**
Ticker	The Company's ticker symbol
Tickers to Compare	Company's ticker symbol for comparison
Model	A drop-down list for choosing different SA models
Parameters	A multi-select list for choosing covariates for SA models

Table 4.1: Description of Input Boxes for Delisting Analysis Interface



(a) Ticker input and Ticker to Compare



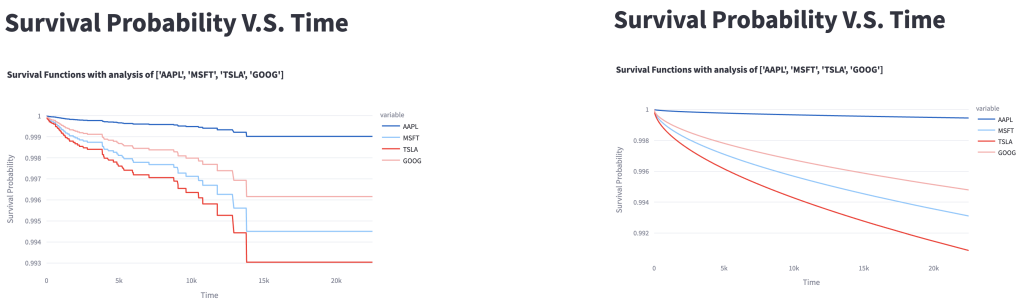
(b) Multi-selection for customer covarients

Figure 4.2: Delisting analysis Side Bar

The Risk of Delisting analysis integrated semi-parametric and parametric analysis as introduced in the section3.3. The 'Risk of Delisting Analysis' presents a sophisticated toolkit within the project, demonstrating a multifaceted approach to survival analysis, particularly tailored for datasets laden with multiple covariates. User can choose different combinations of TVCs calculated as mentioned in the section3.2.1

Central to its design philosophy is the user's freedom to choose between two modelling paradigms: semi-parametric and parametric, as different results are shown in figure4.3. The figure 4.3a shows companies' `AAPL, MSFT, TSLA, GOOG` survival probability against time, considering the user-selected TVCs as covariates. Even though all companies have a very high probability of surviving over 20k days, considering the financial situation of TSLA, it shows a lower survival probability than other companies.

While the semi-parametric models offer flexibility in modelling hazard functions without assuming a specific form, the parametric models, on the other hand, operate under specified distributional assumptions, often leading to more precise estimates given those assumptions hold true. The figure 4.3b shows the result of applying Gamma Fitter to the delisting dataset. The curve is more smooth, giving potentially more accurate predictions of survival probability at different times. This dual approach provides a broadened scope for accuracy, enabling users to tailor their analysis based on the nature of their data and the underlying assumptions they are comfortable making.



(a) The result of Cox PH model

(b) The result of Gamma Fitter

Figure 4.3: Delisting analysis results

After taking the input as described in the table 4.1from the sidebar menu, this project will provide a function to check the correlation between different covariates by displaying a correlation matrix as shown in the figure4.4. Ideally, the user should select covariates with low correlation to others in order to meet the PH assumption. The matrix makes it easy for users to check the correlation between covariates. The whiter, the lower correlation.

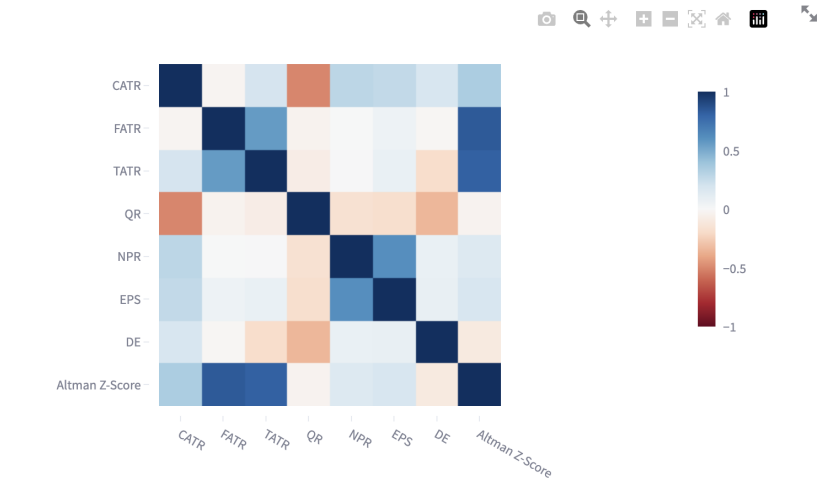
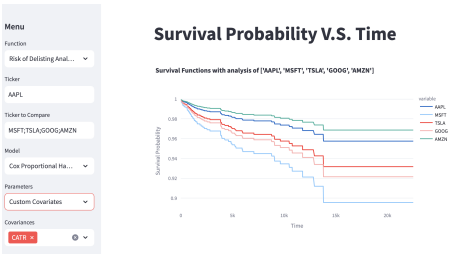
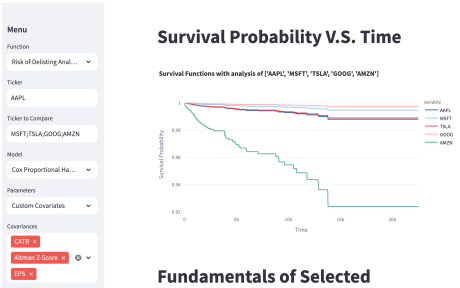


Figure 4.4: The Correlation Matrix

By selecting different companies to compare and different SA model covariates, the new result will be displayed in real-time, enabling analysts to adjust the model according to their needs and aid their investment. As shown in the figure4.5, for the same companies `AAPLl; MSFT; TSLA; GOOG; AMZN`, different combinations of covariates can return different results of the survival curve. Consider TVC `CATR` will indicate the `MSFT` has the lowest risk of survival against time, which the conclusion is not suitable when considering the combinations of three covariates `EPS, Altman Z-Score, CATR` as shown in the figure 4.5b. This time `AMZN` shows the lowest survival curve against time while other companies maintain a high lever of the probability of survival.



(a) The result of Cox PH model with 1 covariate



(b) The result of Cox PH model with 3 covariates

Figure 4.5: Different results of Different numbers of Covariates

4.2.3 The Value at Risk Analysis

As this project defines in the section3.2.2, the VaR analysis represents a repeating event analysis. The model will take in the ticker name, start date, and end date as inputs, shown in the figureA.3. Non-parametric models will be utilized for recurring event analysis, enabling direct derivation of the hazard function and survival curve from the training data without considering the distribution of survival time. The table4.2 shows the description of the input.

Input Box Name	Description
Ticker	The Company's ticker symbol
Start Date	Start Date e.g. 2000-01-01
End Date	End Date e.g. 2023-01-01
Model	A drop-down list for choosing different SA models

Table 4.2: Description of Input Boxes for VaR Analysis

This platform predominantly utilizes non-parametric survival analysis models to delve deeply into the risk of stock returns falling below the VaR (Value at Risk) threshold. The allure of non-parametric methods lies in their ability not to make any a priori assumptions about the probability distribution of the data, thus offering a more precise estimation of the inherent uncertainty and diversity in financial data.

The Kaplan-Meier estimator is among the most widely used within this category. Its core functionality lies in estimating the likelihood of survival over specified time intervals. The data is divided into distinct time segments, with each segment given a survival estimate. This allows observers to clearly gauge the likelihood of stock returns falling below the VaR within a specific time frame.

In contrast, the Nelson-Aalen estimator offers a different perspective. It focuses on calculating the cumulative hazard or the accumulation of risk over a specified time. This gives analysts a more macroscopic risk assessment perspective, especially relevant to long-term investment strategies.

Both Kaplan-Meier and Nelson-Aalen estimators and other models are implemented together in this project, allowing users to select according to their needs. Additionally, since the Risk of Breach VaR is focusing on a single stock, this project also developed functions to display the stock price with its trend and momentum indicators, all displayed on the webpage. The screenshots are attached in the appendixA.2

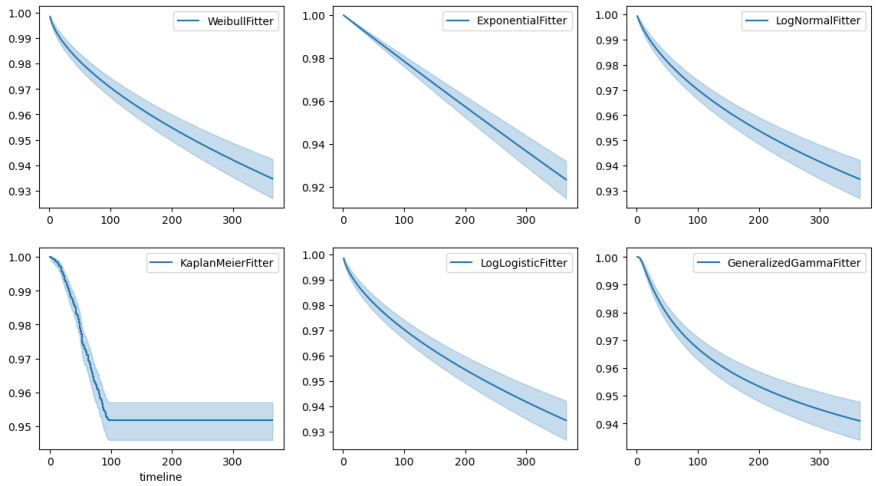


Figure 4.6: The SA Model Results of Breach VaR in 3 Months

The figure 4.6 shows the collection survival function of different fitters, analyzing the risk of a VaR breach event for the AAPL stock. Even though different fitters give roughly the same survival rate at a given time, the different assumptions of the survival time distribution can change the survival curve's shape as the user can visually compare the difference in results between the K-M estimator and Weibull Fitter. This project will not compare the accuracy of different fitters since, according to different conditions, it is difficult to compare the preference of models.

When interpreting results from these models, users can delve into the specific shapes of the survival curves and trends of cumulative hazards. These shapes and trends often encapsulate deeper market insights, such as market sentiments, macroeconomic trends, or other external influencing factors. By analyzing these insights holistically, users can understand the risks associated with stock returns falling below VaR, leading to more informed investment decisions.

Chapter 5

Conclusions

5.1 Summary

The project begins with an exploration of the significance of risk management in the stock market in **Chapter 1**. This importance has been accentuated over the years, especially after catastrophic financial events like the 2008 crisis. Such episodes underscored the dire consequences of inadequate risk management measures. The chapter elucidates the primary objective of the research: to develop a web interface that combines survival analysis with the ever-changing landscape of the stock market. Despite the proven efficacy of survival analysis in various sectors, its application within the stock market remains relatively untapped. This research aspires to bridge this gap.

In **Chapter 2**, the focus shifts to the types of Survival Analysis (SA) models. It delves into the intricacies of various SA models and how they can be effectively applied to analyze behaviours and patterns in the stock market.

Chapter 3 provides a comprehensive breakdown of the methodology adopted for the study. It details the datasets selected, preprocessing methods used, and the models employed, such as the Parametric Regression Model, Cox proportional hazard model, and Non-Parametric Models.

Chapter 4 introduces the design and implementation of the web interface in a detailed introduction to every input for different SA models and explains the results the user will receive.

The project embarked on a multifaceted exploration of the stock market, employing survival analysis as its primary tool. Here's a detailed breakdown of its core activities:

1. **Interface Development**: A cornerstone of this endeavour was the creation of a sophisticated web interface. This interface aimed to seamlessly blend the

principles of survival analysis with the dynamic ebb and flow of stock market trends, offering users an innovative tool to navigate market intricacies.

2. **Delving into Survival Analysis (SA)**: The research dove deep into the family of SA models. It wasn't merely a theoretical exploration; this project actively probed how these models, with their varied characteristics, could be practically applied to dissect stock market behaviours.

3. **Data Handling**:

   - *Collection*: The project sourced specific datasets that held the potential to unlock insights into the stock market's heartbeat.
   - *Preprocessing*: Before any analysis, this raw data underwent rigorous preprocessing to ensure its relevance and accuracy for subsequent modelling efforts.

4. **Tool Creation**: Beyond theoretical analysis, the project bore fruit in the form of a practical tool. This application accepted vital inputs, such as the ticker name and specific date ranges, offering users a tailored analysis of stock market trends.

5. **Analysis of Recurring Events**: One of the project's crowning achievements was its mastery over non-parametric models, specifically tailored for recurring event analysis in the stock market. The project could directly deduce hazard functions and survival curves through these models, bypassing the need for survival time distribution assumptions.

5.2 Limitation and Future Work

5.2.1 Limitations

1. **Scope of Survival Analysis Models**: While the research explored various Survival Analysis (SA) models, the application to the stock market is still in its nascent stage. The complexity and multifractal nature of stock prices might demand more specialized models, for instance, a self-defined linear formula used in the Cox PH model.

2. **Data Limitations**: The study's conclusions are bound by the datasets utilized. There might be inherent biases or missing data points that could affect the gener-

alizability of the results. Also, yfinace can only provide the recent three years' fundamentals or income statements for delisting companies, which can be a challenge for data collection.

3. **Recurring Event Analysis**: While non-parametric models were employed for recurring event analysis, the financial market's unpredictable nature could lead to anomalies that these models may not capture efficiently. For instance, the model is less sensitive to stock return's sudden drop.

4. **Unknown Distribution of Survival Time**: The exact distribution of survival time might be unknown for survival data, making it difficult to determine the most suitable distribution for modelling.

5.2.2 Further Work

1. **Expanding Model Horizon**: Future research can dive deeper into more specialized or hybrid SA models to cater to the unique dynamics of the stock market better.

2. **Incorporating Broader Data**: Integrating data from diverse sources or global stock markets can provide a more holistic view and improve the robustness of the findings.

3. **Enhancing the Tool**: The web interface can be enriched with features that account for global financial news, sentiment analysis, or even AI-driven predictions to make it more versatile.

4. **Real-time Analysis**: Given the dynamic nature of stock markets, real-time analysis and predictions using survival analysis can be a promising avenue for future work.

Bibliography

[1] James Lam. *Enterprise risk management: from incentives to controls*. John Wiley & Sons, 2014.

[2] Mike Adu-Gyamfi. The analysis of the collapse of lehman brothers. *Available at SSRN 2771615*, 2015.

[3] In Lee and Yong Jae Shin. Fintech: Ecosystem, business models, investment decisions, and challenges. *Business horizons*, 61(1):35–46, 2018.

[4] Philippe Jorion. *Value at risk: the new benchmark for managing financial risk*. The McGraw-Hill Companies, Inc., 2007.

[5] René M Stulz. Rethinking risk management. In *Corporate Risk Management*, pages 87–120. Columbia University Press, 2008.

[6] Edward J Sullivan and Timothy M Weithers. Louis bachelier: The father of modern option pricing theory. *The Journal of Economic Education*, 22(2):165–171, 1991.

[7] Epaminondas Panas. Estimating fractal dimension using stable distributions and exploring long memory through arfima models in athens stock exchange. *Applied Financial Economics*, 11(4):395–402, 2001.

[8] Stefan Mittnik and Svetlozar T Rachev. Modeling asset returns with alternative stable distributions. *Econometric reviews*, 12(3):261–330, 1993.

[9] Paul Lévy. Les lois de probabilité dans les ensembles abstraits. *Revue de Métaphysique et de Morale*, 32(2):149–174, 1925.

[10] Young Shin Kim, Svetlozar T Rachev, Michele Leonardo Bianchi, and Frank J Fabozzi. A new tempered stable distribution and its application to finance. In *Risk Assessment: Decisions in Banking and Finance*, pages 77–109. Springer, 2009.

[11] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[12] Adrian Gepp and Kuldeep Kumar. The role of survival analysis in financial distress prediction. *International research journal of finance and economics*, 16(16):13–34, 2008.

[13] Eugene F Fama. Efficient market hypothesis. *Diss. PhD Thesis, Ph. D. dissertation*, 1960.

[14] Eugene F Fama. Random walks in stock market prices. *Financial analysts journal*, 51(1):75–80, 1995.

[15] Antti Ilmanen. Stock-bond correlations. *The Journal of Fixed Income*, 13(2):55, 2003.

[16] Dedy Dwi Prastyo, Titis Miranti, and Nur Iriawan. Survival analysis of companies' delisting time in indonesian stock exchange using bayesian multiple-period logit approach. *Malaysian Journal of Fundamental and Applied Sciences*, 13(4-1):425–429, 2017.

[17] Michele De Laurentiis and Peter M Ravdin. Survival analysis of censored data: neural network analysis detection of complex interactions between variables. *Breast cancer research and treatment*, 32:113–118, 1994.

[18] John P Klein and Prem Goel. Survival analysis: state of the art. 1992.

[19] Arthur V Peterson Jr. Expressing the kaplan-meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association*, 72(360a):854–858, 1977.

[20] David G Kleinbaum, Mitchel Klein, David G Kleinbaum, and Mitchel Klein. Kaplan-meier survival curves and the log-rank test. *Survival analysis: a self-learning text*, pages 55–96, 2012.

[21] Jong Gun Lee, Sue Moon, and Kavé Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomputing*, 76(1):134–145, 2012.

[22] Søren Johansen. An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, pages 165–174, 1983.

[23] Magdalena Constantin and S Das Sarma. Volatility, persistence, and survival in financial markets. *Physical Review E*, 72(5):051106, 2005.

[24] Leonidas Sandoval Junior. Survivability and centrality measures for networks of financial market indices. *arXiv e-prints*, pages arXiv–1201, 2012.

[25] Gregor Wergen. Modeling record-breaking stock prices. *Physica A: Statistical Mechanics and its Applications*, 396:114–133, 2014.

[26] Enrico Scalas 7, Rudolf Gorenflo, Hugh Luckock, Francesco Mainardi, Maurizio Mantelli, and Marco Raberto. Anomalous waiting times in high-frequency financial data. *Quantitative Finance*, 4(6):695–702, 2004.

[27] Suzanne GM Fifield, David M Power, and C Donald Sinclair. An analysis of trading strategies in eleven european stock markets. *The European Journal of Finance*, 11(6):531–548, 2005.

[28] Safwan Mohd Nor and Guneratne Wickremasinghe. The profitability of macd and rsi trading rules in the australian stock market. *Investment Management and Financial Innovations*, (11, Iss. 4 (contin.)):194–199, 2014.

[29] Simon A Levin and Andrew W Lo. Introduction to pnas special issue on evolutionary models of financial markets. *Proceedings of the National Academy of Sciences*, 118(26):e2104800118, 2021.

[30] David Aronson. *Evidence-based technical analysis: applying the scientific method and statistical inference to trading signals*. John Wiley & Sons, 2011.

[31] Charles D Kirkpatrick II and Julie A Dahlquist. *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.

[32] Yakup Kara, Melek Acar Boyacioglu, and Ömer Kaan Baykan. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319, 2011.

[33] Hyejung Chung and Kyung-shik Shin. Genetic algorithm-optimized long short-term memory network for stock market prediction. *Sustainability*, 10(10):3765, 2018.

[34] Eda Orhun and Blanka Grubjesic. Value at risk (var) method: An application for swedish national pension funds (ap1, ap2, ap3) by using parametric model, 2007.

[35] Stefan Klößner and Sven Wagner. Exploring all var orderings for calculating spillovers? yes, we can!—a note on diebold and yilmaz (2009). *Journal of Applied Econometrics*, 29(1):172–179, 2014.

[36] Mazin AM Al Janabi. Equity trading risk management: the case of casablanca stock exchange. *International Journal of Risk Assessment and Management*, 7(4):535–568, 2007.

[37] Joshua Rosenbaum and Joshua Pearl. *Investment banking: valuation, LBOs, M&A, and IPOs*. John Wiley & Sons, 2021.

[38] Ciaran Walsh. *Key management ratios: the 100+ ratios every manager needs to know*. Pearson Education, 2008.

[39] Martin S Fridson and Fernando Alvarez. *Financial statement analysis: a practitioner's guide*. John Wiley & Sons, 2022.

[40] Muhammad Arief Abbas et al. Analisis prediksi tingkat kebangkrutan dengan metode altman z-score pada perusahaan sub sektor makanan dan minuman yang listing di bursa efek indonesia periode 2010-2016. *Jurnal Manajemen Update*, 7(1), 1983.

[41] Rihfenti Ernayani. Predicting the potential bankruptcy of coal mining companies using altman z-score method during 2012-2016 period. *Humanities & Social Sciences Reviews*, 8(1):491–500, 2020.

[42] CMA Melwani and Manish Sitlani. Study of financial performance and its determinants: Empirical evidence from listed indian 2/3 wheeler manufacturer firms. In *Proceedings of 10th International Conference on Digital Strategies for Organizational Success*, 2019.

[43] M Swalih, K Adarsh, and M Sulphey. A study on the financial soundness of indian automobile industries using altman z-score. *Accounting*, 7(2):295–298, 2021.

[44] Dan Hauschild. Altman z-score: Not just for bankruptcy. *From Z-score to" Green Zone" survivability: AMPros Corporation*, 2013.

[45] Md Sohel Rana, Sk Masum Billah, Mohammed Moinuddin, Md Abu Bakkar Siddique, and Md Mobarak Hossain Khan. Exploring the factors contributing to increase in facility child births in bangladesh between 2004 and 2017–2018. *Heliyon*, 9(5), 2023.

[46] Johannes Bock. Quantifying macroeconomic expectations in stock markets using google trends. *arXiv preprint arXiv:1805.00268*, 2018.

[47] yfinance · pypi. https://pypi.org/project/yfinance/. (Accessed on 08/09/2023).

[48] Api documentation — alpha vantage. https://www.alphavantage.co/documentation/. (Accessed on 08/09/2023).

[49] Qingli Dong, Yingwei Peng, and Peizhi Li. Time to delisted status for listed firms in chinese stock markets: An analysis using a mixture cure model with time-varying covariates. *Journal of the Operational Research Society*, 73(10):2358–2369, 2022.

[50] Testing the proportional hazard assumptions — lifelines 0.27.7 documentation. https://lifelines.readthedocs.io/en/latest/jupyter$_notebooks/Proportional(Accessedon08/15/202$3

[51] Mats J Stensrud and Miguel A Hernán. Why test for proportional hazards? *Jama*, 323(14):1401–1402, 2020.

[52] Streamlit documentation. https://docs.streamlit.io/. (Accessed on 08/13/2023).

Appendix A

First appendix

A.1 The PseudoCode for MontCarlo Simulation

A.2 The SA Model Results for AAPL and Trend, Momentum Indicators

A.3 The Side Bar of Breach VaR Analysis

Algorithm 1 VaR Calculation using Monte Carlo Simulation

Import required libraries (pandas, numpy, datetime, scipy)
procedure LOADDATA
    Read "aapl_data.csv" into a dataframe called *aapl_data*
    Convert the 'Date' column to datetime format
    Sort the dataframe by 'Date' in ascending order and set 'Date' as the index
end procedure
procedure COMPUTEDAILYRETURNS
    Calculate the daily returns using the 'Close' column
    Store the result in a new column named 'Returns'
end procedure
procedure MONTECARLOVAR(start_date, stock_data, forecast_days=63, simulations=100000)
    Extract daily returns mean and standard deviation up to *start_date*
    Initialize a matrix for simulated stock prices
    for each day in forecasting window **do**
        Simulate stock prices based on daily returns mean and standard deviation
    end for
    Calculate the 5% VaR using simulated stock prices
    return VaR_95
end procedure
procedure DETECTVARBREACH
    Create an empty list named *VaR_event_records*
    for each rolling window in data **do**
        Set the *start_date* as the start of forecasting window
        Calculate VaR using MONTECARLOVAR
        for each day in forecasting window **do**
            if 'Close' price breaches VaR **then**
                Record the breach date and exit loop
            end if
        end for
        Add breach details to *VaR_event_records*
    end for
end procedure
procedure SAVERESULTSTODATAFRAME
    Convert *VaR_event_records* to dataframe *VaR_event_df*
    Sort *VaR_event_df* by 'Prediction_Start_Date'
end procedure
Display VaR_event_df

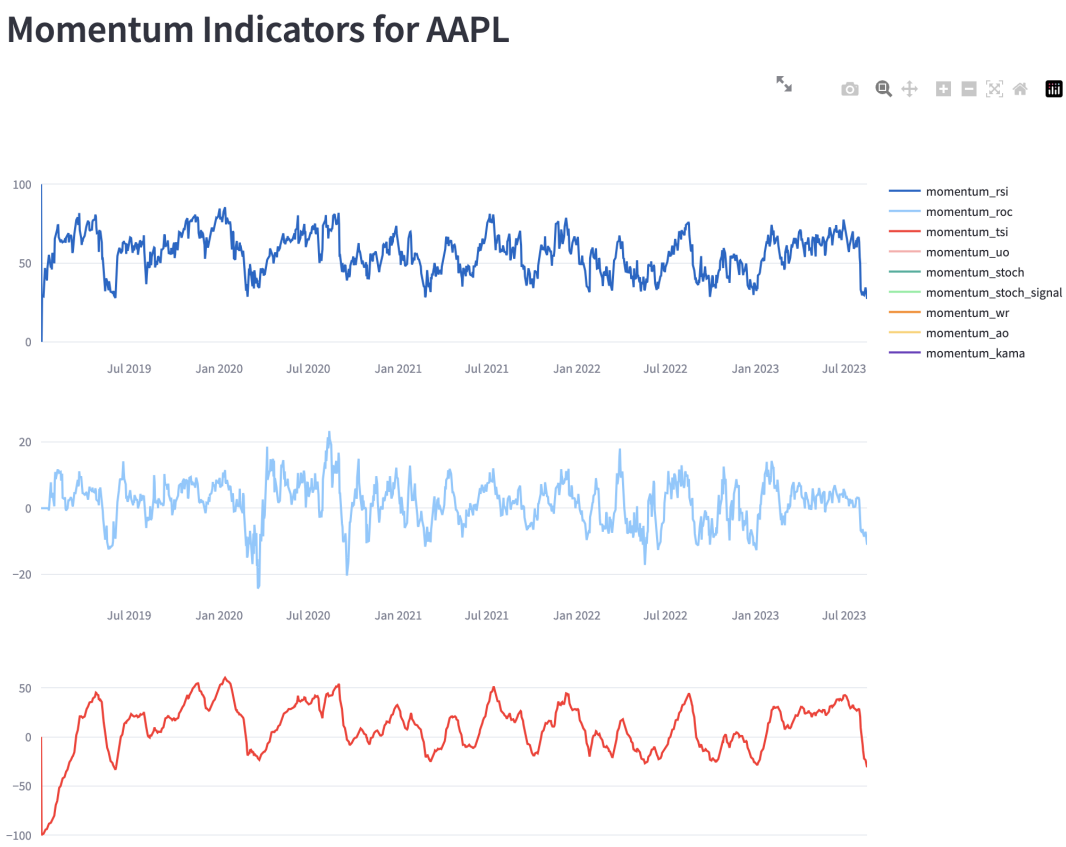
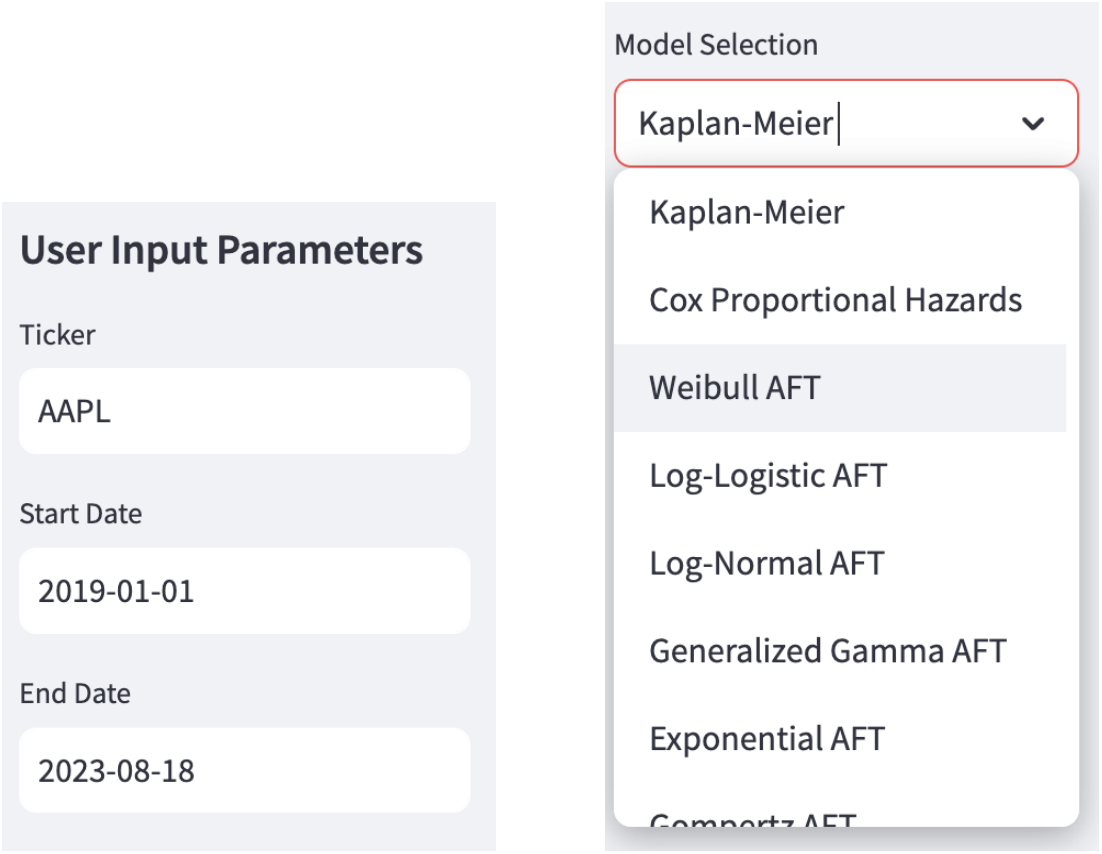


Figure A.1: Momentum Indicators for AAPL

Figure A.2: Trends for AAPL



(a) Ticker Input and Date Input

(b) The Model Selection Menu

Figure A.3: The Side Bar