The University of Hong Kong

Faculty of Engineering

Department of Computer Science

COMP7705

Project Report

Systematic Risk Warning Function of Financial Derivatives

Submitted in partial fulfillment of the requirements for the admission to the degree of Master of Science in Computer Science

By

Li Jiayao (3036032516)

Tang Yutian (3036033821)

Xia Linlong (3036034837)

Li Dongheng (3036034875)

Supervisor: Dr. Zhang Jingrui

Date ofsubmission: 01/08/2023

Abstract

With the continuous development of China's capital market and the continuous improvement of residents' income in recent years, more and more investors begin to enter the capital market, and capital market investment has certain risks. In order to protect the interests of investors and prevent and resolve systemic financial risks, stock market risk early warning is a key measure. Among them, financial derivatives are closely related to preventing and resolving systemic risks in the stock market. In order to investigate the early warning function of financial derivatives on the stock market risk in China, this paper constructs explanatory variables based on the price, trading volume, trading amount and open position of CSI 300 stock index futures, predicts the stock market risk based on the machine learning models, and compares the early warning effect of derivatives of different varieties under the circumstances of weekly frequency, monthly frequency and quarterly frequency. The results show that financial derivatives have early-warning function for stock market risk, and the short-term early-warning function is better than the long-term one. This project has enriched the research of stock market risk early warning and provided enlightenment for using financial derivatives to improve stock market early warning ability.

Key words: Stock market risk; Financial derivatives; Supervised machine learning

Declaration

We, Li Jiayao, Tang Yutian, Xia Linlong and Li Dongheng, declare that this report titled " Systematic Risk Warning Function of Financial Derivatives" and the work presented herein are entirely our original work, except where otherwise indicated. We affirm that no part of this report has been submitted for publication elsewhere or has been plagiarized from any other sources.

We further declare that any research material, figures, tables, or data used in this report from other sources are duly acknowledged with proper citations and references. Any contributions made by individuals or organizations to this research are appropriately recognized.

We hereby state that this research has not received any funding or sponsorship that could create a potential conflict of interest in relation to the content and findings presented in this report.

The data and materials used in this study are available upon request, and we are willing to provide any additional information to interested parties.

Li Jiayao, Tang Yutian, Xia Linlong and Li Dongheng

Department of Computer Science, University of Hong Kong

August 1, 2023

Table of contents

I. Introduction

The global financial markets are highly complex and interconnected, often exhibiting significant volatility and risk. Financial derivatives have emerged as essential tools for market participants to manage and hedge against these risks. These derivatives provide investors with the opportunity to speculate, hedge, and diversify their portfolios[1]. In recent years, there has been an increasing interest in understanding the relationship between derivatives and their respective spot markets, particularly in terms of their risk warning effect[1]. The risk warning effect refers to the ability of derivatives to act as early indicators or signals of potential downward movements in the underlying spot markets. Identifying and analyzing these warning signals can be crucial for investors, traders, and policymakers in making appropriate decisions and managing their exposure to market risks.

Financial derivatives have revolutionized the way investors and financial institutions manage market risks. These instruments derive their value from underlying assets, such as stocks, bonds, commodities, or currencies. Notable examples include futures, options, swaps, and forwards. By providing a means to transfer and diversify risk, derivatives play a pivotal role in mitigating market uncertainties. Investors can use derivatives to protect their portfolios from adverse price movements, thereby hedging against potential losses. Furthermore, derivatives offer opportunities for speculation, enabling investors to capitalize on market trends and generate returns in both bullish and bearish market conditions.

Understanding the risk warning effect of financial derivatives is of paramount importance in financial markets. As derivatives are often tied to the performance of their underlying assets, they can provide valuable insights into potential market

movements. For instance, the price of stock index futures can be indicative of the expected future performance of the underlying stock market index. A significant increase or decrease in the price of futures may signal potential bullish or bearish sentiments in the spot market, respectively. Analyzing such early warning signals can empower investors to make informed decisions, adjust their investment strategies, and manage risk exposure proactively.

Supervised machine learning models have already garnered significant attention in recent years due to their capability to learn patterns and relationships from historical data and make accurate predictions. At the same time, machine learning algorithms have seen extensive application in the field of finance, revolutionizing the way financial institutions analyze data, make predictions, and manage risks. One notable area where machine learning has gained significant traction is in quantitative trading and investment strategies. Financial firms now employ sophisticated machine learning models to analyze vast amounts of historical market data, seeking patterns and trends that can inform investment decisions and optimize trading strategies. These algorithms can identify market inefficiencies, predict price movements, and dynamically adjust trading positions, resulting in more efficient and profitable trading strategies. Machine learning has also found applications in credit risk assessment and fraud detection. Financial institutions use machine learning models to analyze customer data, transaction patterns, and credit history to assess creditworthiness accurately. By automating the credit risk assessment process, lenders can make faster and more reliable credit decisions, improving customer experience and reducing credit risk exposure. Furthermore, machine learning is transforming financial forecasting and market analysis. By processing a wide range of macroeconomic indicators and financial data, these algorithms can predict market trends, interest rates, and economic conditions with

greater accuracy. This enables financial professionals to make informed decisions and allocate resources more effectively.

Overall, the recent applications of machine learning algorithms in finance have demonstrated their potential to enhance decision-making processes, improve risk management strategies, and optimize investment performance, making them indispensable tools for financial institutions in an increasingly data-driven and competitive landscape.

These models are trained using labeled datasets, where input variables (features) are paired with corresponding output values (labels). By leveraging supervised machine learning, we can construct predictive models that discern patterns in market data and forecast future market movements.

For our research, we plan to employ various supervised machine learning algorithms, such as linear regression, decision trees, random forests, and support vector machines, etc, to analyze the relationship between financial derivatives and their risk warning effect on their corresponding spot markets. We will train these models using historical market data in China’s stock exchange, including price, trading volume, and inventory of derivatives and their corresponding spot assets, and add the control variables which are actually significant in the risk identification in the capital market in China. By validating the models' performance on test datasets, we can assess their predictive accuracy and suitability for forecasting market risks.

To conduct our research, we will utilize extensive market data from reputable financial databases, including iFind and CSMAR, etc. These databases provide a vast array of market information, encompassing multiple asset classes, maturities, and frequencies. Our dataset spans from April 2010 to December 2022, making the time frame of our

research as longer as possible, ensuring a comprehensive view of market dynamics over a significant period.

The research will focus on CSI 300 stock index futures, a prominent derivative instrument in the Chinese market. By leveraging this dataset, we will investigate the risk warning effect of these futures on the underlying spot market i.e, CSI 300 stock index. We will analyze the predictive performance of different supervised machine learning models and compare the early warning abilities of derivatives with varying frequencies, such as weekly, monthly, and quarterly scenarios.

Effective risk management is a cornerstone of successful financial market participation. Investors, traders, and financial institutions face various risks, including market risk, credit risk, liquidity risk, and operational risk. These risks can have adverse effects on portfolio performance, financial stability, and overall market health. Implementing robust risk management strategies is essential to protect investments and navigate through turbulent market conditions. By exploring the risk warning effect of financial derivatives, our research aims to contribute to the advancement of risk management practices in global financial markets. By identifying early warning signals and understanding the relationship between derivatives and spot markets, investors can make better-informed decisions and adjust their investment strategies in response to changing market conditions.

In summary, our research project seeks to shed light on the risk warning effect of financial derivatives on global financial markets. Leveraging supervised machine learning algorithms and utilizing a comprehensive dataset, we aim to provide practical implications for market participants and contribute to the existing body of knowledge in the field of risk management. By understanding the predictive abilities of derivatives and their relationship with spot markets, investors can enhance their risk management

practices and navigate the complexities of financial markets with greater confidence. As financial markets continue to evolve, the insights gained from our research will remain relevant for investors, traders, and policymakers seeking to optimize risk management strategies and secure financial well-being.

II. Literature Review

There are many scholars have made their efforts in the research of the impact of financial derivatives on the corresponding spot market as well as implementing the machine learning methods to explore the risk management tasks in financial industry.

2.1 Literatures about the Impact of Financial Derivatives on the Corresponding Spot Market

Financial derivatives play a crucial role in price discovery, enabling efficient anticipation of stock market risk. They provide a mechanism for market participants to assess and incorporate new information into securities prices as well as derivatives. Research conducted by Fleming et al. (1996) explored the dynamics of rational market which is perfectly frictionless in the same time, concluding that the newly appeared information in the market must be reflected by the prices of securities and corresponding derivatives simultaneously. Interestingly, they found that stock index futures had a guiding effect on the spot market, indicating their significance in influencing market behavior [2].

Another noteworthy study by An et al. (2004) delved into the correlation between the implied volatility of options and the future returns of corresponding underlying stock. The researchers discovered that stocks exhibiting significant increment in implied volatilities of their call (put) options over the previous month tended to have high (low) stock returns in the future, particularly for options with high trading volumes.

These finding sheds light on the potential predictive ability of implied volatility of options in forecasting future stock market movements [3].

Furthermore, Pan and Poteshman (2006) uncovered a significant relationship between trading volume of options and the corresponding underlying future stock prices. By leveraging a unique dataset, they constructed put-call ratios as indicators based on option volume initiated by buyers when opening new positions. Intriguingly, stocks with high put-call ratios underperformed those with low put-call ratios by more than 40 basis points on the subsequent day and over 1% during the following week. This suggests that monitoring put-call ratios can offer valuable insights into future stock performance [4].

Additionally, Xing et al. (2010) explored the predictive power of the volatility smirk, which refers to the shape of the volatility curve of put options. They found that the skewness of the volatility curve of put options had a significant cross-sectional correlation with future equity returns. Specifically, a greater skewness corresponded to lower future stock returns. This observation highlights the potential usefulness of the volatility smirk as an indicator for assessing future market movements [5].

Apart from facilitating price discovery and risk anticipation, financial derivatives also play a role in hedging systemic risk. The volatility index (VIX), which is constructed by taking the statistics of implied volatility of options, has proven to be an effective predictor of systemic risk. Jiang and Tian (2005) conducted a study using options on the Standard & Poor's 500 (SPX) index and found that VIX reliably predicted future realized volatility. Their results indicated that all the information embedded in the Black-Scholes implied volatility and past realized volatility are included in the implied volatility, allowing it to predict future volatility more effectively[6].

Furthermore, Pan et al. (2019) demonstrated the value of incorporating VIX in volatility prediction models, such as GARCH models. Their research highlighted the significant ability of VIX to predict stock volatility. Moreover, they observed that VIX information greatly reduced option pricing errors, underscoring its usefulness in enhancing pricing accuracy [7].

Bohl, Salm, and Schuppli (2011) found that the existing literature on price discovery in stock index futures and spot markets has often overlooked the influence of distinct investor groups. In their study, they address the gap by examining the time-varying linkages between spot and futures markets using a VECM-DCC-GARCH framework, while also considering changes in the investor structure of the futures market over time. their empirical findings indicate that when the futures market is primarily dominated by presumably uninformed private investors, it does not actively contribute to price discovery. However, as institutional investors' share in trading volume rises, they observed evidence of information transmission from futures to spot markets and a notable increase in conditional correlation between both markets. [8]

In Xie and Huang (2014)’s study, they utilize daily data of the China Securities Index (CSI) 300 spanning from 2005 to 2012 to examine the influence of index futures trading on spot market volatility in China. By employing a series of GARCH models, they derive several key findings: The introduction of index futures does not lead to a decrease in the volatility of the spot market. They observe a reduction in sensitivity to new information following the launch of the CSI 300 index futures, while sensitivity to historical information increases. Lastly, their analysis does not reveal any evidence of a leverage effect either before or after the introduction of the CSI 300 index futures. These findings contribute to a better understanding of the relationship between index futures trading and spot market volatility in the Chinese market. [9]

Xu and Zhang (2023) have made their efforts in research into the high-frequency CSI300 data. They focused on predicting the high-frequency one-minute CSI300 first distant futures trading volume, a crucial aspect for market participants and policymakers. The research utilizes neural network and finds that rading volume can be accurately predicted using data from one to thirty minutes ahead, with a relatively low-complexity model featuring five hidden neurons. Incorporating nearby futures or spot trading volumes generally does not improve predictions significantly. The results have implications for designing trading platforms, assessing system risk, and forming index price predictions, making them valuable to both market practitioners and policymakers. [10]

Ausloos, Zhang, & Dhesi (2020) advocates the use of TGARCH modeling to investigate the impact of index futures trading on spot price variability, with a focus on the CSI-300 index as a test case. The findings reveal that the introduction of CSI-300 index futures trading leads to a significant reduction in volatility in the spot market. Additionally, a stationary equilibrium relationship is observed between the CSI-300 spot and CSI-300 index futures markets, and bidirectional Granger causality is detected. Furthermore, the research concludes that spot prices can be predicted with higher accuracy using a time span of 3 or 4 lag days. [11]

Bohl, Diesteldorf, & Siklos (2015) ‘s study investigates the impact of introducing Chinese stock index futures on the volatility of the underlying spot market. Utilizing Generalized Auto-regressive Conditional Heteroscedasticity (GARCH) models, the researchers compare findings for mainland China with Chinese index futures traded in Singapore and Hong Kong. The results reveal that Chinese index futures lead to a decrease in spot market volatility in all three markets examined. However, this effect is not observed in the companion index futures markets in Hong Kong and Singapore.

Despite China's relatively young and retail investor-dominated stock market, the evidence supports the stabilization hypothesis, typically observed in mature markets. [12]

2.2 Literatures about Machine Learning Methods to Explore the Risk Management Tasks in Financial Industry

Behera, Pasayat and Kumar (2023) explore a novel portfolio construction technique that combines machine learning algorithms for stock return prediction with a mean-VaR (value-at-risk) model for portfolio selection. The future performance of stock markets is considered a crucial factor in portfolio creation, and with the advancements in machine learning techniques, new opportunities arise for incorporating prediction concepts into the selection process. The hybrid approach involves utilizing various machine learning regression models, including Random Forest, Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Support Vector Machine Regression (SVR), k-Nearest Neighbors (KNN), and Artificial Neural Network (ANN), to forecast stock values for the next period. The findings indicate that the mean-VaR model with AdaBoost prediction outperforms the other models, demonstrating its effectiveness in constructing portfolios with improved performance. [13]

Niu, Wang and Zhang (2023) focus on examining the influence of geopolitical risks on predicting US stock market volatility using machine learning models. The empirical findings reveal that military build-ups and escalation of war have significant importance in predicting realized volatility among various geopolitical risks. The research emphasizes the superior performance of machine learning and forecast combination methods, particularly the SVR method and trimmed mean combination. Moreover, the study demonstrates that by allocating a portfolio based on machine

learning-based volatility forecasts, particularly using elastic net and random forest, a mean-variance investor can achieve substantial financial benefits. [14]

Chen, Wen and Nanehkaran, etc (2023) highlight the significance of stock market analysis for investors in managing the risk and achieving profit. Their research focuses on two crucial tasks in stock price analysis: predicting stock prices and identifying graphic signals from candlestick charts. The experimental results demonstrate that the proposed ML-based approaches outperform the state-of-the-art methods, showing superior performance. The methods are found to be reliable and applicable in real-world stock exchange strategies. [15]

2.3 Research Gap and Objective

By summarizing the previous relevant literatures, we find that: Although there have been extensive studies on the systemic risk of the stock market and financial derivatives, few have examined the role of financial derivatives as stock market risk warnings. Previous research in this area has primarily focused on macro and stock market perspectives, yielding suboptimal results. For instance, Bussiere and Fratzscher (2006) achieved only 42.3% signal accuracy when using various indicators from international and domestic economies to warn of stock market risks [16].

To address these limitations, it becomes imperative to explore alternative approaches. Traditional measurement methods often struggle to process large volumes of data efficiently. In contrast, machine learning algorithms have shown promise in improving predictive power, particularly when dealing with high-dimensional variables. Hence, our research aims to assess the effectiveness of financial derivative indicators in enhancing stock market risk warnings and explore the advantages of

employing machine learning algorithms to improve the predictive accuracy of risk warning models.

III. Methodology

3.1 Variables Settings

3.1.1 Explained Variable: Stock Market Risk Factor

Stock market risk usually act as a sharp decline in the stock index, and we use the ratio of the current period (closing) index to the maximum value of that index over the past k+1 period (period t-k to period t) to indicate the degree of stock market decline, defined as the maximum loss ratio:

$$\mathrm{ML_t} = \frac{\mathrm{P_t}}{\max(\mathrm{P_{t-k}}, \mathrm{P_{t-k+1}}, \ldots, \mathrm{P_t})} \qquad (1)$$

$\mathrm{P_t}$ denotes the closing stock index level on the last trading day of the t-th week (month, quarter) and k is the number of lags. In the following, the lag period is uniformly set to 1 year. Therefore, k = 52 for weekly frequency forecasting, and k equals 12 and 4 for monthly and quarterly frequency forecasting, respectively.

On the basis of the ML series calculated from equation (1), we further calculate the market risk coefficient (CC) :

$$\mathrm{CC}_t = \begin{cases} 1 & if\ ML_t < \overline{ML_t} - 1.5\sigma_t \\ 0 & others \end{cases} \qquad (2)$$

$\overline{\mathrm{ML_t}}$, σ_t are respectively the mean, standard deviation of the ML series in the past k periods (from period $\mathrm{t-k}$ to t-1 period).

From equation (2), the risk factor CC divides the ML series into two categories: 1 for the occurrence of an overall stock market price decline event and 0 for the non-occurrence of an overall stock market price decline event.

3.1.2 Main Explanatory Variables

We choose the price, position and other indicators of derivatives to portray the information of derivatives market. Firstly, returns on stock index futures are chosen as explanatory variables for stock market risk forecasting. Secondly, the trading volume, amount and inventory of derivatives. These variables imply information on the extent of investor disagreement on future price expectations of the derivatives market. Intuitively, this information is closely related to and, in turn, has an impact on the price jumps in the spot market. Therefore, a direct study of the relationship between trading volume, amount and inventory of derivatives and stock market risk is also necessary.

The names, symbols and calculation methods of these indicators are shown in the table:

Table 1. Definition of explanatory variables

Name	Symbol	Definition
Futures Returns	$\Delta \ln (\mathrm{FCLS}_t)$	The difference of natural logarithm of the ratio of closing price of futures contracts in period t over the closing price of it in period t-1. If the indicator is constructed using weekly frequency data, then the arithmetic mean of the closing prices of futures contracts for all trading days in week t is first calculated and then take the natural logarithmic difference. (Similar for monthly and quarterly data)
Futures Inventory	$\ln (\mathrm{FOI}_t)$	The natural logarithm of futures contract inventory. If weekly frequency data is used to construct the indicator, then the inventory for all trading days in week t is summed first. (Similar for monthly and quarterly data)
Futures Volume	$\ln (\mathrm{FVOI}_t)$	The natural logarithm of futures contract volume. If weekly frequency data is used to construct the indicator, then the volume for all trading days in week t is summed first. (Similar for monthly and quarterly data)

Name	Symbol	Definition
Futures Amount	$\ln(\mathrm{FAMT}_t)$	The natural logarithm of the trading amount (unit: 100 million yuan) of futures contracts traded. If the indicator is constructed using weekly frequency data, then the amount for all trading days in week t is summed first. (Similar for monthly and quarterly data)
Spot and Futures Basis Difference	BASIS_t	The difference between the closing price of the stock index futures contract and the spot closing index. If weekly frequency data is used to construct the indicator, then the basis is calculated for each trading day, and then the arithmetic mean of all trading days in week t is taken.

3.1.3 Control Variables

In our project, we decide to use the stock market's own factors as control variables in order to investigate whether financial derivatives have the function of enhancing the effectiveness of stock market risk warning. Considering the frequency of variable update and data availability, the psychological line index (PLI), relative strength index (RSI) and investor confidence index (ICI) of the stock market are selected as control variables.

In addition, our study also includes the index (constituent) turnover amount, financing balance as a percentage of control variables. The index trading amount (ITA) can portray the panic behavior of investors in the period of overall price decline in the stock market, so this indicator is used as a control variable for stock market risk prediction. Considering that the financing balance is affected by the turnover amount, and also in order to eliminate the volume, we divide the financing balance of Shanghai and Shenzhen markets by the turnover number of stocks in both markets, i.e., the financing balance ratio to represent the market killing atmosphere. Stock market risk is usually expressed as liquidity depletion. Therefore, the liquidity of the stock market is highly correlated with the risk, so the turnover ratio is used to represent the liquidity, and the

ratio of the turnover of the index components to the total equity in circulation is used to calculate the turnover ratio. Our research also constructs the difference between the 10-year Treasury bond and the 6-month Treasury bond yield as the Treasury spread as control variables.

3.2 Model Construction

Our objective was to create a dependable predictive model for recognizing and evaluating systematic risk warning signals in financial derivatives. To achieve this goal, we plan to develop and assess six machine learning models: ①Support Vector Machine (SVM), ②Naive Bayes, ③Gradient Boosting, ④AdaBoost, ⑤Random Forest, and ⑥Linear Regression. Our aim was to explore various models and determine the most suitable ones for the prediction of systematic risk warning in the stock market.

3.3 Sample Selection and Prediction Process

Firstly, in order to comprehensively examine the predictive effect of financial derivatives on stock market risk. We select the beginning date of the first financial derivative in China, the CSI 300 stock index futures, as the starting point of the series. Therefore, the sample interval is from April 2010 to December 2022.

Table 2. Sample Selection and Data Sources

Name		Sample interval	data source
Daily Frequency Data	CSI 300 Index, CSI 300 Stock Index Futures, and Index Turnover, Financing Balance Ratio, Exchange Rate lot ratio, treasury spread	2010/4/16-2022/12/31 (Interval of CSI 300 index starts at 2008/3/31)	iFind database
Weekly Frequency (Monthly	Mental Line Index (MLI), Relative Strength Index (RSI), Investor Confidence Index (ICI)	2010/4/1 - 2022/12/31	

Frequency Frequency) Data			

After we obtain the combined dataset for CSI 300 (weekly, monthly and quarterly), we will implement some necessary steps before we feed the data into our selected machine models: (1) min-max standardization; (2) data augmentation. Then, we will split our data into training dataset (80%) and test dataset (20%), then feed them into the models to test which one is the most suitable and accurate for the stock risk prediction task.

IV. Data Analysis

4.1 Data Acquisition

All the datasets related to the CSI300 Index and Futures have been acquired from iFind Dataset. The CSI 300 Index is a market capitalization-weighted index that tracks the performance of the top 300 stocks listed on the Shanghai and Shenzhen stock exchanges. The CSI300 Futures are derivative contracts that allow investors to speculate on the future direction of the CSI300 Index.

By acquiring datasets on CSI300 Index Level, CSI300 Future Volume, CSI300 Future Inventory, CSI300 Future Amount, and CSI300 Future Closing Price, further data preprocessing and analysis can be carried out. The acquired datasets cover a substantial time frame, spanning from April 16, 2010, to December 31, 2022, ensuring an extensive and diverse range of information for our analysis.

4.2 Data Preprocessing and Descriptive Statistics

In this section, data preprocessing steps and the descriptive statistics for the explanatory variables, explained variable, and control variables that are used to investigate whether financial derivatives have a risk warning effect on the financial market.

4.2.1 Explanatory Variables

After acquiring the data, the raw data has been transformed into the explanatory variables in order to investigate whether financial derivatives have a risk warning effect on the financial market. We used the acquired datasets to calculate the 5 explanatory variables: Future Returns Ratio $\Delta\ln(\mathrm{FCL}S_t)$, Futures Inventory $\ln(\mathrm{FO}I_t)$, Futures Volume $\ln(\mathrm{FVO}I_t)$, Futures Amount $\ln(\mathrm{FAM}T_t)$, and Spot and Futures basis difference $\mathrm{BASI}S_t$. We have transformed all five variables into weekly, monthly, and quarterly frequencies. By creating a descriptive statistic of these variables across multiple time frequencies, we can gain a preliminary understanding of the potential risk warning effect on the financial market.

We have summarized the statistics result of all five explanatory variables during two periods: the normal period (CC = 1) and the declining period (CC = 0):

Table 3. Descriptive statistics: normal period vs. declining period

Name	Symbol	Mean	Std.	Mean(normal)	Mean(decline)	Significance of mean difference
Futures returns (Monthly)	$\Delta ln(FCLS$	0.0011	0.0582	0.0066	-0.0242	0.0000***
Futures returns (Weekly)		0.0002	0.0279	0.0039	-0.0156	0.0000***
Futures returns (Quarterly)		0.0073	0.1013	0.0342	-0.0756	0.0006***
Spot and Futures basis difference (Monthly)	$BASIS_t$	-8.4633	23.9308	-5.7618	-20.8700	0.0026***

Spot and Futures basis difference (Weekly)		-8.4066	28.9119	-5.1078	-22.7031	0.0000***
Spot and Futures basis difference (Quarterly)		-9.0855	22.6444	-6.6374	-16.6338	0.1868
Futures inventory (monthly)	$ln(FOI)_t$	0.9795	0.0183	0.9800	0.9770	0.4415
Futures inventory(weekly)		12.8897	0.7611	12.9186	12.7584	0.0354**
Futures inventory(quarterly)		15.4826	0.6657	15.4923	15.4527	0.8600
Futures volume(monthly)	$ln(FVOL)$	14.8339	1.4251	14.9394	13.3492	0.0509*
Futures volume(weekly)		13.3507	1.4441	13.3937	13.1711	0.1239
Futures volume (quarterly)		15.9556	1.4472	16.0136	15.7770	0.6277
Futures amount (monthly)	$ln(FAMT)$	10.2340	1.3475	10.3310	9.7882	0.0576*
Futures amount (weekly)		8.7529	1.3702	8.7798	8.6375	0.2998
Futures amount (quarterly)		11.3595	1.3700	11.4032	11.2250	0.6997

Note:

(1)***, ** and * are significant at 1%, 5% and 10% levels respectively.

(2) The mean difference test against the futures inventory, volume, and amount is tested against the raw data arithmetic mean, not the natural logarithm data, because the dimensionality of the data is eliminated by the natural logarithm, as is the numerical difference between the normal interval and the mean during the decline period.

Based on the table, the future returns, Spot, and Futures Basis Differences differ significantly between the normal and declining periods based on the p-value. In the Significance of mean difference column, for instance, "0.000***" indicates a highly significant difference, "0.035**" suggests a moderately significant difference, "0.090*" implies a slightly significant difference, and values greater than 0.1 generally indicate non-significant differences.

Therefore, based on the provided chart, the following variables are significant in the two periods:

(1) Future Returns Ratio (Monthly):

The mean of Future Returns Ratio (Monthly) during the "normal" period is 0.0066, indicating a positive average monthly return during normal periods.

The mean of Future Returns Ratio (Monthly) during the "declining" period is -0.0242, indicating a negative average monthly return during declining periods.

The "Significance of mean difference" is denoted as "0.0000***," which means that the difference in means between the "normal" and "declining" periods is highly statistically significant. The p-value is less than 0.001 (indicated by three asterisks), indicating strong evidence against the null hypothesis of no difference in means.

(2) Future Returns Ratio (Weekly):

The mean of Future Returns Ratio (Weekly) during the "normal" period is 0.0039, indicating a positive average weekly return during normal periods.

The mean of Future Returns Ratio (Weekly) during the "declining" period is -0.0156, indicating a negative average weekly return during declining periods.

The "Significance of mean difference" is denoted as "0.0000***," which means that the difference in means between the "normal" and "declining" periods is highly statistically significant. The p-value is less than 0.001, indicating strong evidence against the null hypothesis of no difference in means.

(3) Future Returns Ratio (Quarterly):

The mean of Future Returns Ratio (Quarterly) during the "normal" period is 0.0342, indicating a positive average quarterly return during normal periods.

The mean of Future Returns Ratio (Quarterly) during the "declining" period is -0.0756, indicating a negative average quarterly return during declining periods.

The "Significance of mean difference" is denoted as "0.0006***," which means that the difference in means between the "normal" and "declining" periods is statistically significant at a significance level of 0.001 (indicated by three asterisks).

(4) Spot and Futures Basis Difference (Monthly):

The mean of Spot and Futures Basis Difference (Monthly) during the "normal" period is -5.7618.

The mean of Spot and Futures Basis Difference (Monthly) during the "declining" period is -20.8700.

The "Significance of mean difference" is denoted as "0.0026***," which means that the difference in means between the "normal" and "declining" periods is statistically significant at a significance level of 0.01 (indicated by three asterisks).

(5) Spot and Futures Basis Difference (Weekly):

The mean of Spot and Futures Basis Difference (Weekly) during the "normal" period is -5.1078.

The mean of Spot and Futures Basis Difference (Weekly) during the "declining" period is -22.7031.

The "Significance of mean difference" is denoted as "0.0000***," which means that the difference in means between the "normal" and "declining" periods is highly statistically significant (p-value < 0.001).

(6) Spot and Futures Basis Difference (Quarterly):

The mean of Spot and Futures Basis Difference (Quarterly) during the "normal" period is -6.6374.

The mean of Spot and Futures Basis Difference (Quarterly) during the "declining" period is -16.6338.

The "Significance of mean difference" is denoted as "0.1868," which means that the difference in means between the "normal" and "declining" periods is not statistically significant at the conventional significance levels (p-value > 0.05).

Looking at the provided table, some of the variables for different frequencies have p-values less than or equal to 0.05, indicating statistical significance, while others have p-values greater than 0.05, indicating no statistical significance.

The significance of a variable's mean difference can depend on several factors, such as the sample size, variability of data, and the actual differences in means between the two periods. Variables with larger sample sizes and more substantial differences in means are more likely to yield smaller p-values and, hence, show statistical significance.

It's important to interpret the significance of mean differences in the context of the specific data and the underlying research question. Non-significant results do not necessarily imply that the variables are not relevant; they may simply suggest that

there is not enough evidence in the data to conclude that the means are significantly different between the two periods.

To provide a more specific explanation for each variable's lack of significance, a detailed analysis of the data, hypothesis testing methodology, and sample sizes would be required. So we will continue to conduct EDA towards those variables in “4.3 Exploratory Data Analysis”

4.2.2 Explained Variables

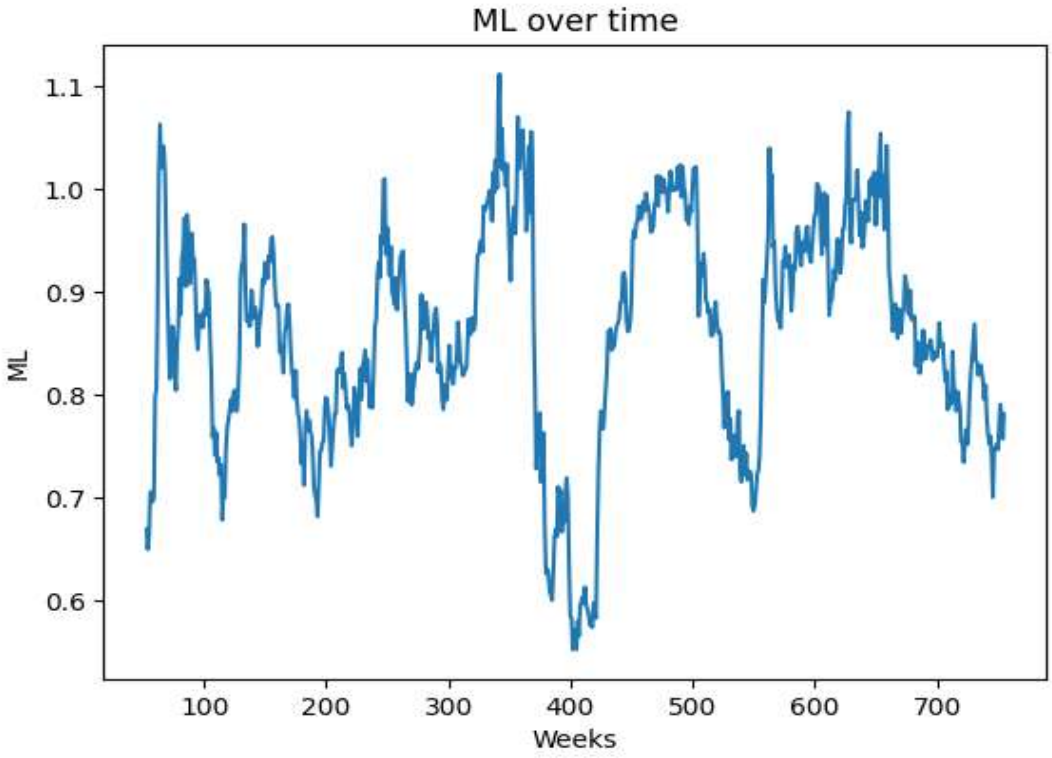

Figure 1. Weekly ML trend

Then we calculate the explained variables, Maximum Loss (ML), and risk coefficient (CC) using the two formulas provided in the Methodology section into weekly, monthly, and quarterly frequencies. The ML (Maximum Loss) over time graph indicates that the ML value dropped to nearly 0.5 from week 400 to week 420 (around September 2017

to June 2018). This suggests that the financial stock market has experienced a decline during this time period. As we know, the supply-side reform in 2018 drove up the cost of midstream manufacturing; Deleveraging leads to the expansion of enterprises into debt crisis, listed companies into equity pledge crisis; The trade war between China and the United States has dealt a heavy blow to exporters and caused great uncertainty in the market, leading to a major bear market in the stock market in 2018.

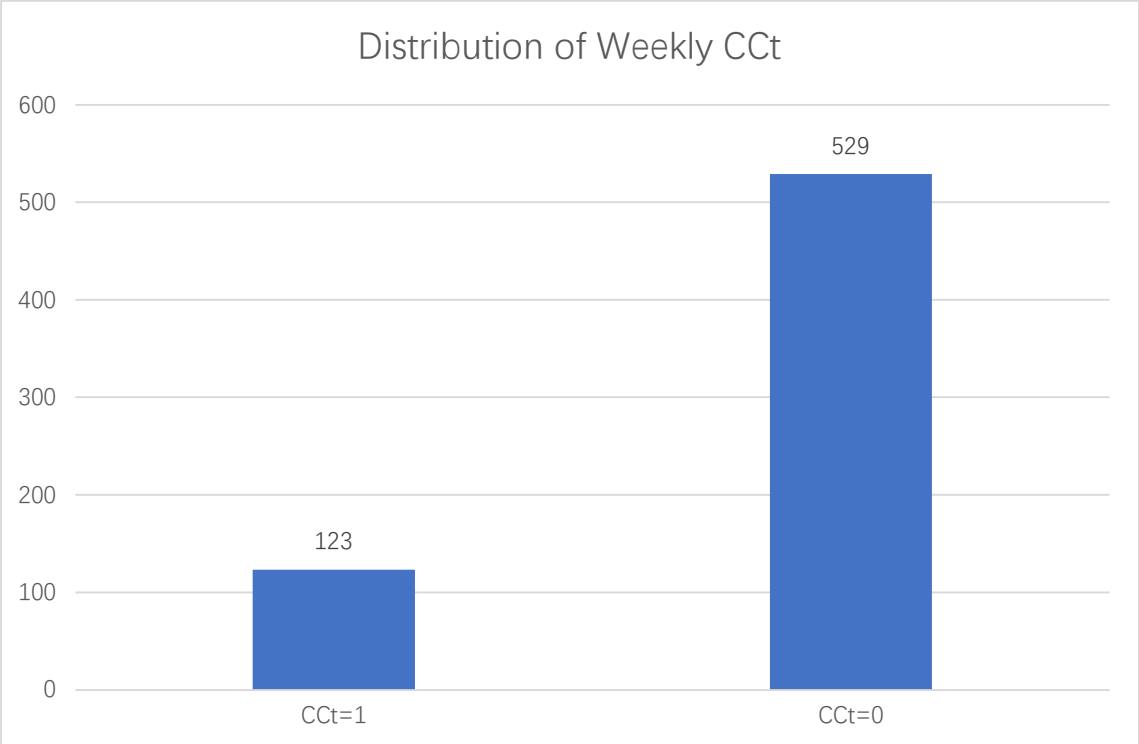

Figure 2. Weekly CCt distribution

However, most of the time the ML value stays above 0.7, indicating that ML exhibits a certain degree of "clustering", and stock market crashes are not common. According to the CC's (Risk Coefficient) formula (2) in the Methodology section, the general decline in the stock market prices occurs when CC equals 1, and CC equals 0 represents the absence of a general decline in stock market price. For our dataset, we have 123 decline week occurrences (18.87% of the total sample period) according to the CC distribution graph.

4.2.3 Control Variables

As stated in the Methodology section, six control variables are created. We have also summarized the statistics result of all six control variables during two periods: the normal period (CC = 1) and the declining period (CC = 0); and each variable will be grouped by weekly, monthly, and quarterly:

Table 4. Descriptive statistics of Control Variables: normal period vs. declining period

CSI 300-weekly-control variables					
Name	**Mean**	**Std.**	**Mean(normal)**	**Mean(decline)**	**Significance of mean difference**
Index_turnover_rate	0.5159	0.3425	0.5311	0.4355	0.0090***
short_balance_ratio	0.034	0.0472	0,0348	0.0291	0.2532
pli	0.5172	0.2326	0.5212	0.4965	0.3213
t_spread	0.7919	0.3652	0.7943	0.7791	0.696
RSI	51.4185	25.2567	51.3162	51.9572	0.8126
index_trading_volume	8.6839	0.7806	8.6998	8.5996	0.2299
CSI 300-monthly-control variables					
Name	**Mean**	**Std.**	**Mean(normal)**	**Mean(decline)**	**Significance of mean difference**
Index_turnover_rate	0.5165	0.3237	0.5298	0.4547	0.2755

short_balance_ratio	0.0331	0.0465	0.0317	0.0398	0.4146
pli	0.5111	0.1101	0.515	0.4927	0.3424
t_spread	0.7952	0.3602	0.7794	0.8686	0.2447
RSI	50.4442	9.3741	50.0506	52.2662	0.2668
index_trading_volume	10.1552	0.7325	10.1454	10.2005	0.7243
CSI 300-quarterly-control variables					
Name	**Mean**	**Std.**	**Mean(normal)**	**Mean(decline)**	**Significance of mean difference**
Index_turnover_rate	0.5188	0.3031	0.5341	0.4754	0.5534
short_balance_ratio	0.032	0.0453	0.0308	0.0353	0.7599
pli	0.5125	0.0692	0.5364	0.4443	0.000***
t_spread	0.7923	0.3275	0.746	0.9242	0.0917*
RSI	49.7578	5.7894	51.0913	45.9625	0.0048***
index_trading_volume	11.2737	0.7087	11.2714	11.2802	0.9698

For the weekly analysis, the mean values of index turnover rate, PLI, and RSI for the CSI 300 index are slightly higher during the normal period compared to the declining period. The difference in mean values of index turnover rate between the normal period and decline period is statistically significant (0.009***). This suggests that the index turnover rate is higher in the normal period compared to the decline period.

For the monthly analysis, the mean value of index turnover rate, PLI are higher in the normal period compared to the decline period; and the mean value of other control variables in monthly analysis are higher in the decline period. Furthermore, the mean

difference between the two periods of all control variables is not statistically significant in the monthly analysis.

For the quarterly analysis, the mean of PLI, RSI, index turnover rate is higher in the normal period; and PLI and RSI are significantly higher during the normal period (0.000*** and 0.0048***).

Overall, the descriptive statistics present an overview of the control variables of CSI300 Index. The analysis shows that control variables such as PLI and RSI have a significant mean difference between the normal period and decline period. This suggests that investor sentiment is an important factor that affects the stock market's performance; and PLI and RSI will be categorized as key variables in the model training and evaluation section later on.

4.3 Exploratory Data Analysis

In this section ,we will use the weekly dataset as the example to show the distribution and correlation

4.3.1 Data Cleaning

Data cleaning was the initial step in the procedure, where we removed any errors in the data, thereby ensuring data accuracy and completeness. "Not a Number" (NaN) entries were removed, including the first and last rows of the processed CSI300 dataset, where NaN entries were present.

By removing the NaN entries, the dataset is ready for subsequent analysis. Other critical steps in data cleaning, such as filtering unwanted outliers and handling missing data, are typically necessary, they were not required in this case. This is because the data source was authentic and all the variables were preprocessed before

integration into the dataset, ensuring that the dataset was already free from unwanted outliers and missing data.

After the completion of the data cleaning procedure, the processed dataset contains 650 entries and 13 features including the label. All variables are represented as float values.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 650 entries, 1 to 650
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   week                  650 non-null    int64
 1   future_return         650 non-null    float64
 2   volume                650 non-null    float64
 3   inventory             650 non-null    float64
 4   amount                650 non-null    float64
 5   BASIS                 650 non-null    float64
 6   index_turnover_rate   650 non-null    float64
 7   short_balance_ratio   650 non-null    float64
 8   pli                   650 non-null    float64
 9   t_spread              650 non-null    float64
 10  RSI                   650 non-null    float64
 11  index_trading_volume  650 non-null    float64
 12  CCt+1                 650 non-null    float64
dtypes: float64(12), int64(1)
memory usage: 71.1 KB
```

Figure 3. Variables Snapshot

4.3.2 Univariate Analysis

Univariate analysis involves analyzing a single variable at a time to examine its individual distribution and characteristics. All the variables will be analyzed, and key variables identified in the univariate analysis are typically used in both bivariate analysis and multivariate analysis.

Label CC_{t+1} has a total of 650 entries, with 527 entries that is equal to 0 and 123 entries that is equal to 1. Variable CC_{t+1} is skewed towards the 0 value; and only small number of entries that is equal to 1. The percentage of major decrease in market price event is only 18.87% in the processed CSI300 dataset. This imbalance in the distribution of the label may be a significant issue in the future model training and evaluation, since the small number of observations with a value of 1 may result in reduced statistical power.

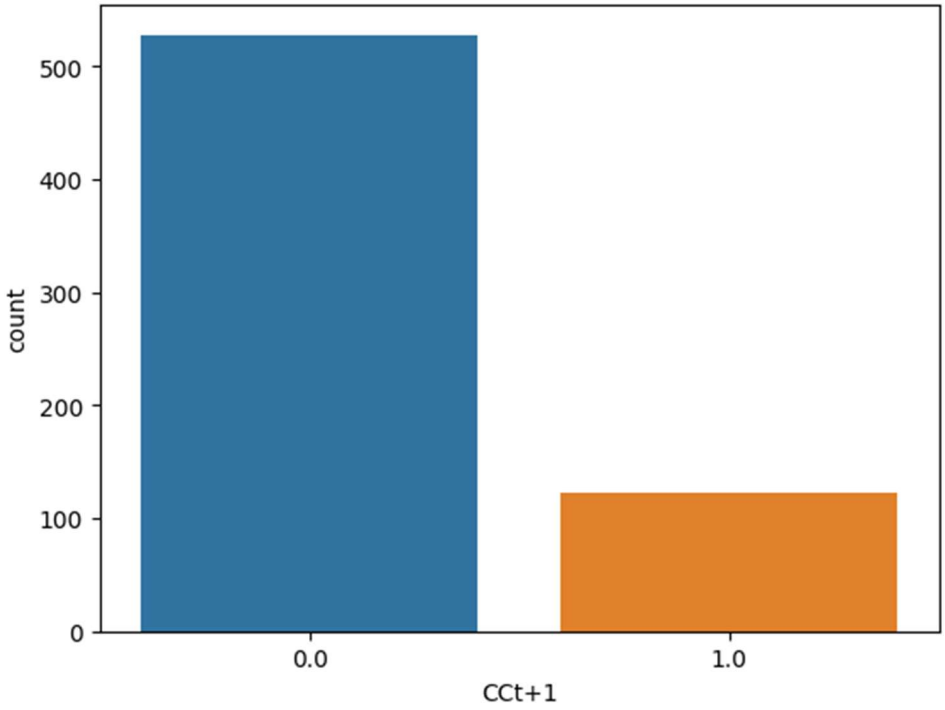

Figure 4. CCt+1 Distribution

To address the issue of class imbalance, oversampling methods will be considered, such as synthetic minority oversampling technique (SMOTE), to ensure that the selected model accurately captures the relationship between the variable and the response. SMOTE can generate synthetic samples by interpolating between existing

minority class samples. Later in this section, balanced dataset using SMOTE technique will be introduced and analyzed.

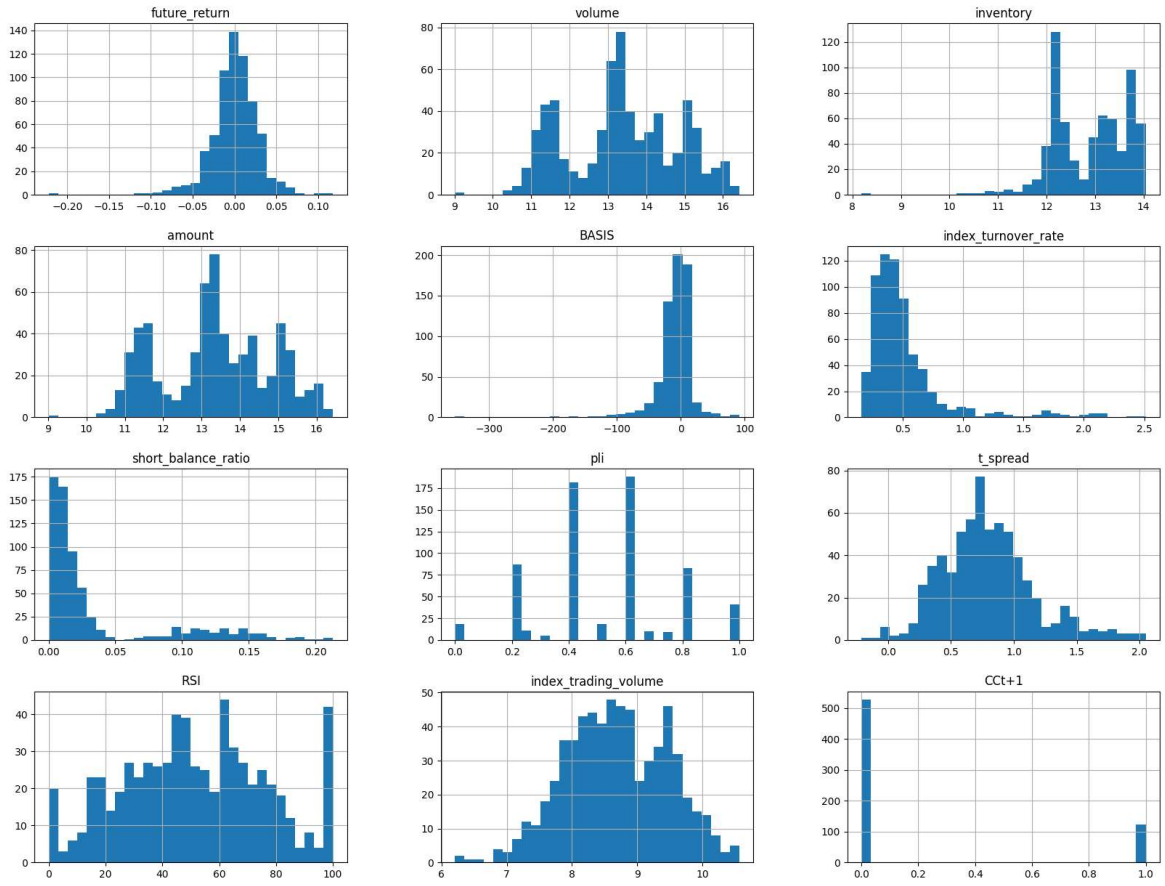

Figure 5. Univariate Histogram

The above histogram group depicts the distribution of all variables in the CSI300 dataset. By examining the shape of the histogram, it has been observed that outliers are present in variables such as inventory, short balance ratio, and index turnover ratio. These outliers may serve as indicators of a potential decrease in the CSI300 market price event.

Furthermore, it is noteworthy to mention that variables such as t-spread, RSI, and index trading volume have similar shapes of distributions, which indicates a strong

relationship between them. Additionally, variables such as index turnover ratio and short balance ratio also present the potential for a strong relationship.

Below is the summary statistics table for all variables:

	future_return	volume	inventory	amount	BASIS	index_turnover_rate	short_balance_ratio	pli	t_spread	RSI	index_trading_volume	CCt+1
count	651.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	652.000000	651.000000
mean	0.000195	13.350687	12.889745	8.752206	-8.402610	0.515855	0.033911	0.517229	0.791938	51.418457	8.683882	0.188940
std	0.027945	1.444134	0.761162	1.369308	28.911996	0.342474	0.047231	0.232607	0.365177	25.256708	0.780547	0.391762
min	-0.222283	9.014325	8.185907	4.415341	-353.150333	0.158000	0.000005	0.000000	-0.210760	0.000000	6.206515	0.000000
25%	-0.012952	12.135342	12.245586	7.497640	-17.671075	0.327500	0.006613	0.400000	0.558505	33.080686	8.132763	0.000000
50%	0.001133	13.331771	13.021572	8.945200	-3.319400	0.424000	0.013465	0.600000	0.748075	50.200947	8.663530	0.000000
75%	0.015916	14.400114	13.606349	9.577401	5.082271	0.562500	0.028256	0.600000	0.984020	69.011539	9.298937	0.000000
max	0.117596	16.422424	14.043919	12.041496	91.457600	2.512000	0.213129	1.000000	2.047400	100.000000	10.568039	1.000000

Figure 6. Statistics Snapshot

These statistics provide a snapshot of the market. For example, the mean RSI is 51.48, which indicates that the market is neither overbought nor oversold. Therefore, the market is at a neutral level. This suggests that there is no clear direction in that specific period of the market and that the buying and selling pressures are relatively balanced. Additionally, the mean index turnover rate is higher than 0.5, indicating that the specific time of the market in the processed CSI300 dataset is relatively active.

4.3.3 Bivariate Analysis

Based on the information provided in the univariate analysis section, some key variables are identified that show promise for further bivariate analysis. These variables are futures inventory, short balance ratio, index turnover ratio, and BASIS.

1. Futures Inventory and Risk Coefficient

In the univariate analysis, the histogram of futures inventory's distribution indicates that it may serve as an indicator of potential market price decreases. Therefore, analyzing its correlation with risk coefficient could yield value insights.

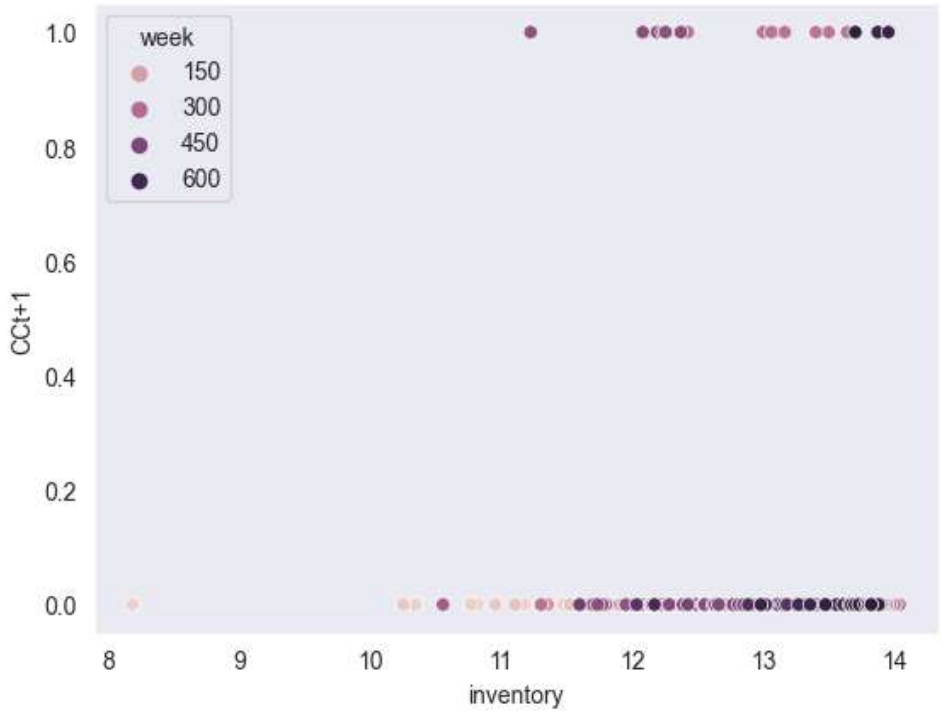

Figure 7. Futures Inventory and Risk Coefficient

The scatterplot indicates that there is no clear linear trend between the futures inventory and risk coefficient. Most of the points that represent major market decrease events, which is when risk coefficient is equal to 1, are clustered towards the upper end of future inventory values. This suggests that higher futures inventory may be a potential indicator of higher market risk. However, there are also many points with higher futures inventory but no market decrease events. This indicates that false positives will occur during the prediction. Therefore, it may need to be combined with other variables to improve performance.

2. Short Balance Ratio and Risk Coefficient

The scatter plot between short balance ratio and risk coefficient was analyzed; and the analysis revealed a weak negative linear trend. This suggested that the risk coefficient tends to be higher at lower short balance ratio values and vice versa.

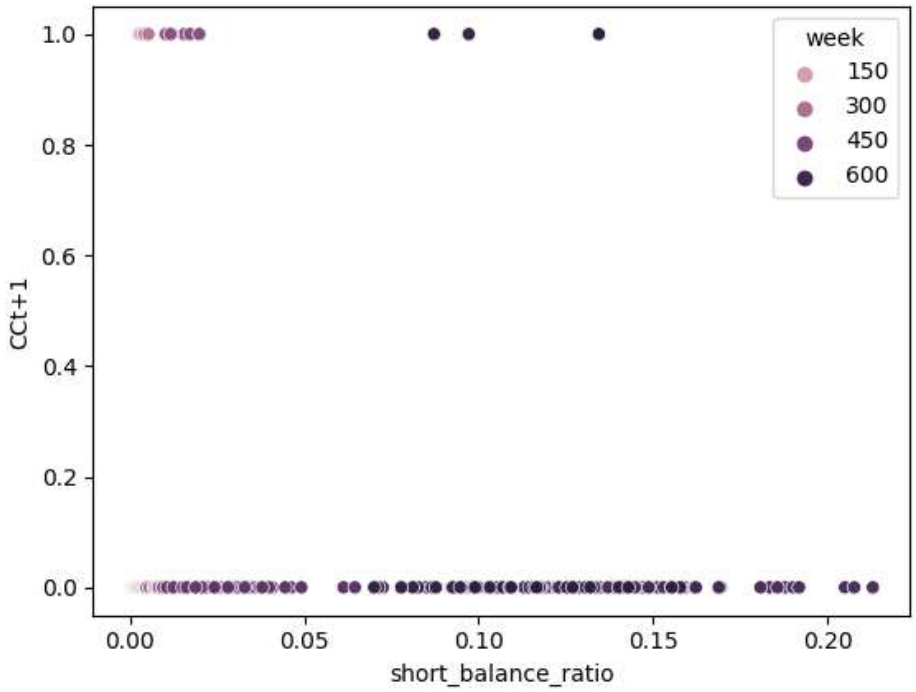

Figure 8. Short Balance Ratio and Risk Coefficient

The analysis also revealed that most points representing major market decrease events were clustered towards the lower end of short balance ratio values. However, there are still many points with lower short balance ratios with no major market decrease events occurring, indicating that the false positive rates will be quite high when using short balance ratio along for prediction.

3. Index Turnover Ratio and Risk Coefficient

Based on the scatterplot between index turnover ratio and risk coefficient, there does not seem to be a clear linear relationship between index turnover ratio and risk coefficient. The points are spread out with no discernible pattern. Most points representing major market decrease event do not cluster around any index turnover ratio values. In this case, both high false positive rate and false negative rate will occur if index turnover ratio is used solely for prediction.

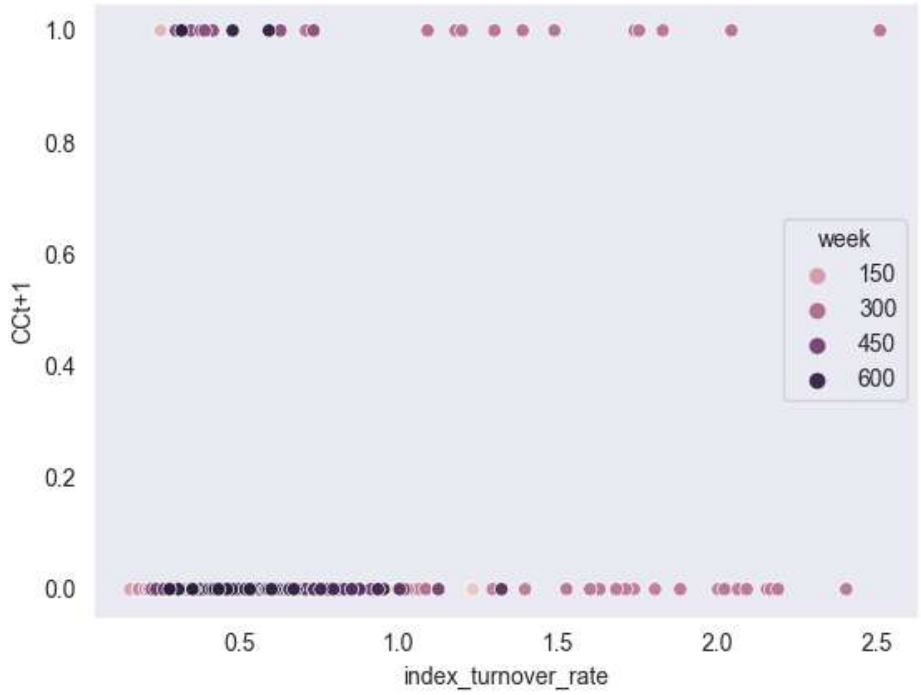

Figure 9. Index Turnover Ratio and Risk Coefficient

The analysis suggests that index turnover ratio alone appears to have limited predictive power for risk coefficient. Therefore, it may need to be combined with other variables to improve performance.

4. Index Turnover Ratio and Short Balance Ratio

The scatter plot between index turnover ratio and short balance ratio does not indicate a clear linear relationship between the two variables. The lack of a distinct linear trend suggests that incorporating index turnover ratio with short balance ratio along may have limited predictive power for the market.

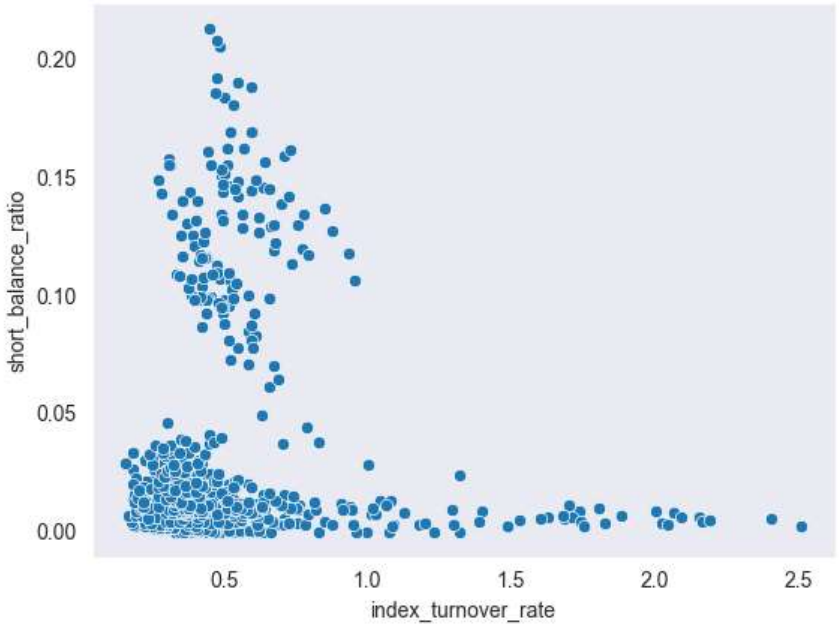

Figure 10. Index Turnover Ratio and Short Balance Ratio

5. BASIS and Risk Coefficient

The scatter plot between BASIS and risk coefficient was analyzed; and the analysis revealed a weak negative linear trend. This suggested that the risk coefficient tends to be higher at lower short balance ratio values and vice versa. Moreover, points representing major market decrease event are all below 0. This indicates that lower BASIS values may be a potential indicator of higher market risk.

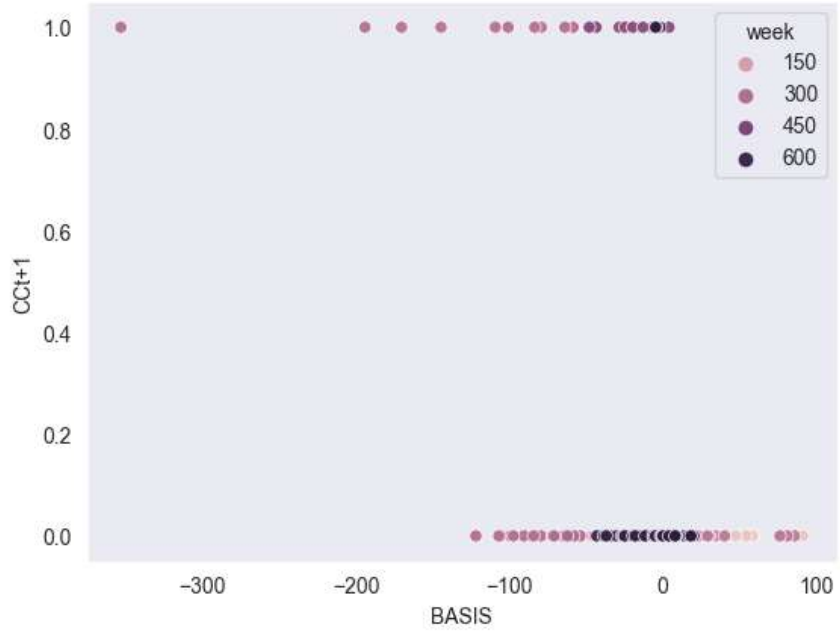

Figure 11. BASIS and Risk Coefficient

However, some points representing low BASIS with no major market price decrease events occur in the dataset. This suggests a high false positive rate possibility when using BASIS alone for major market price decrease event prediction.

4.3.4 Multivariate Analysis

The scatter plot matrix below represents all the variables in the processed CSI300 dataset while considering the impact of CCt+1 on the relationship. From the scatter plot matrix, it is obvious that some variables have a clear liner relationship. For example, there is a strong positive relationship between index trading volume and index turnover ratio. At the same time, some other variables exhibit more random scattering. From example, short balance ratio and inventory appear to have a somewhat linear relationship with more variability.

-

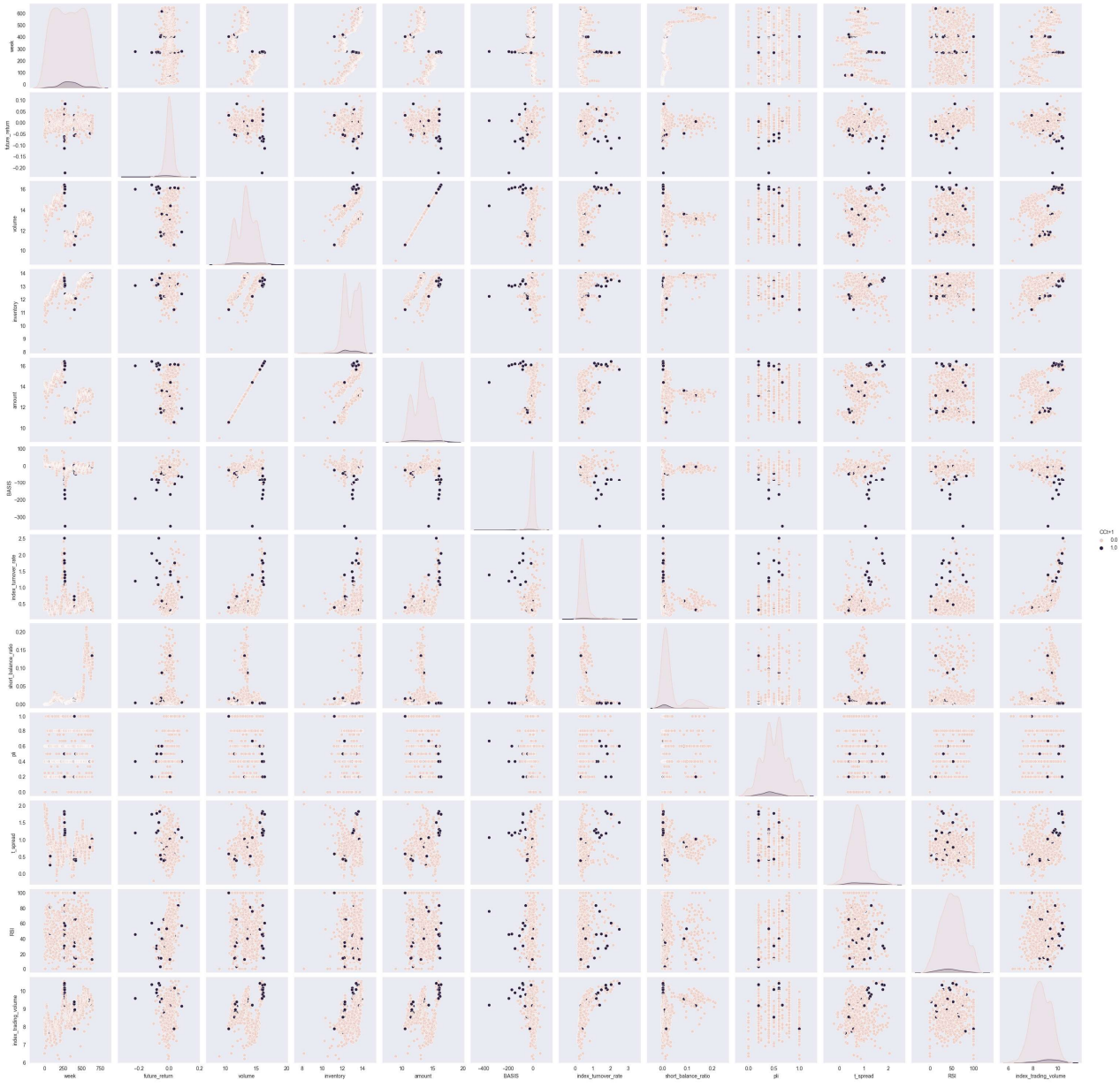

Figure 12. Multivariate Analysis

Using risk coefficient as the hue in each scatterplot provides the opportunity to identify the impact of risk coefficient on the relationship between two variables in the CSI300 dataset. As indicated in the scatter plot matrix, the class imbalance in the distribution of the label (risk coefficient) may affect the process of model training, prediction, and evaluation.

4.3.5 EDA Conclusion

The univariate analysis revealed that the label in our CSI300 dataset is skewed towards 0. This imbalance in the distribution of the label may pose a significant issue in the future model training and evaluation. The bivariate analysis identified some key variables that show promise for further analysis, including futures inventory, short balance ratio, index turnover ratio, and BASIS. Furthermore, the scatter plot matrix in the multivariate analysis provided a comprehensive view of the relationships between all pairs of variables with the impact of risk coefficient.

The findings in the EDA suggest that a combination of variables may be necessary to improve the performance of the selected model in the next step. This may involve incorporating key variables identified in the bivariate analysis as well as others not yet identified. Furthermore, normalization and oversampling techniques such as SMOTE and ROS are required to address the issue of class imbalance problem for the label before model training and evaluation phase. By addressing these issues, the processed CSI300 dataset is ready for the next phase: model training and evaluation.

V. Model Results

After combining all the processed explanatory variables, explained variables and control variables in to 3 combined datasets (weekly, monthly, quarterly) respectively. We have conducted necessary descriptive statics and exploratory data analysis in those combined datasets.

5.1 Data Augmentation

5.1.1 Original Label Distribution

By observing the distribution of our dataset label (CCt), we have discovered that our dataset are imbalanced datasets: that is to say, the situation that the stock market

significant decline happens (CCt = 1) is the minority situation, while the normal periods dominate the dataset at most of the time.

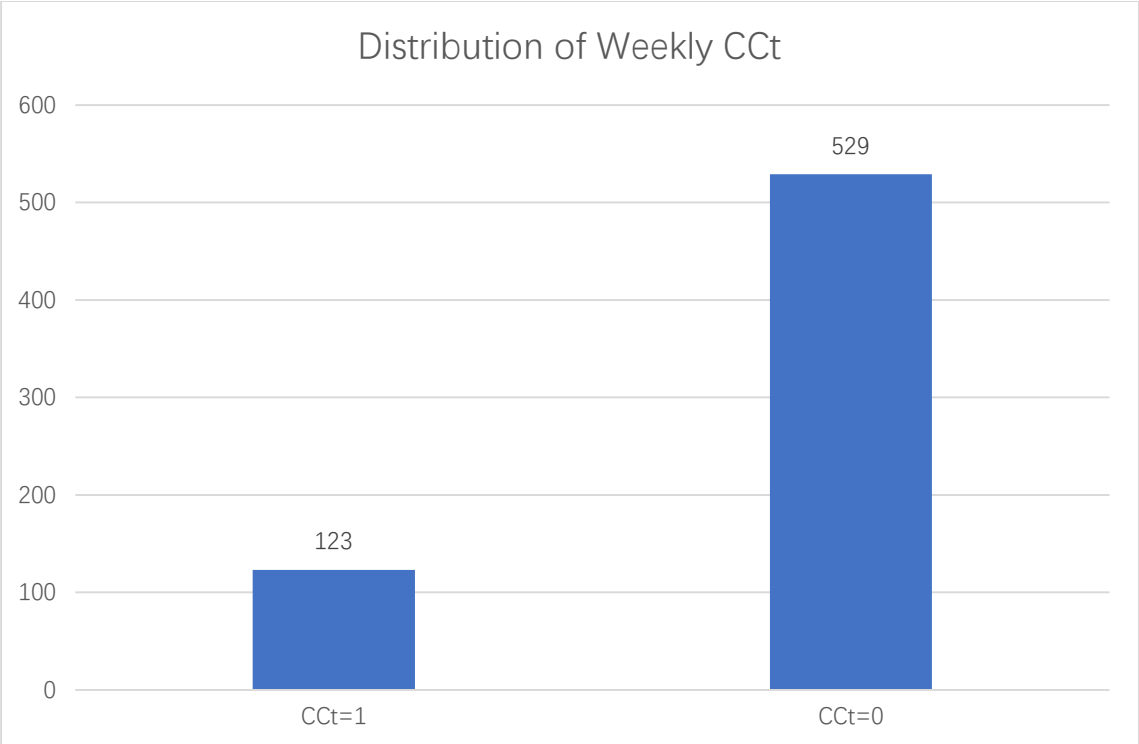

Figure 13. Distribution of Weekly CCt

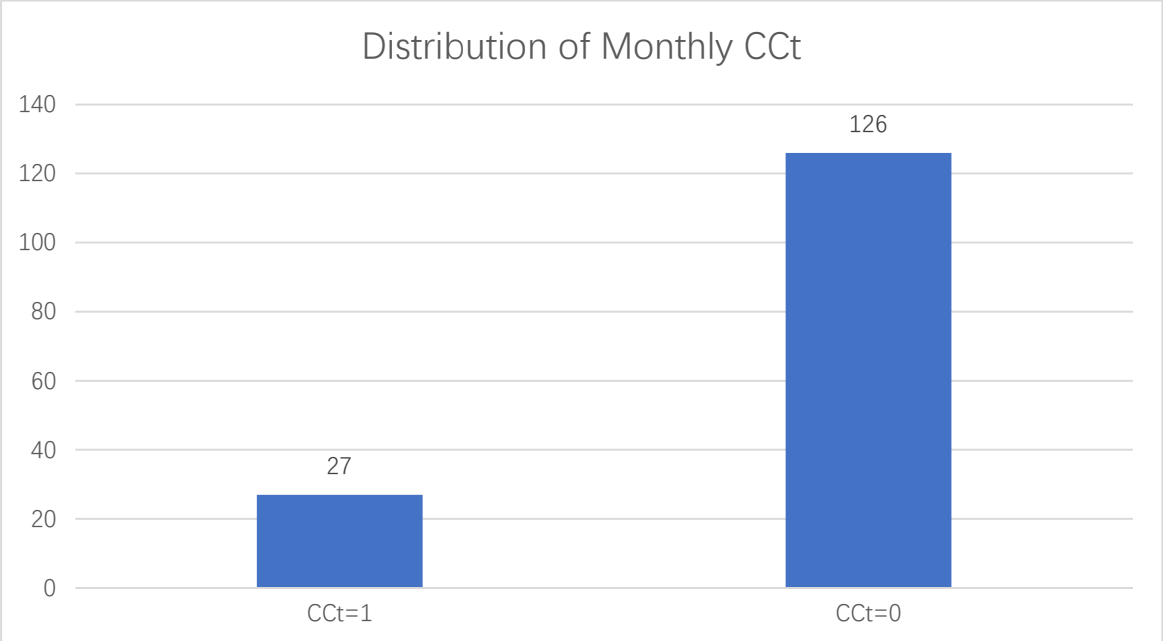

Figure 14. Distribution of Monthly CCt

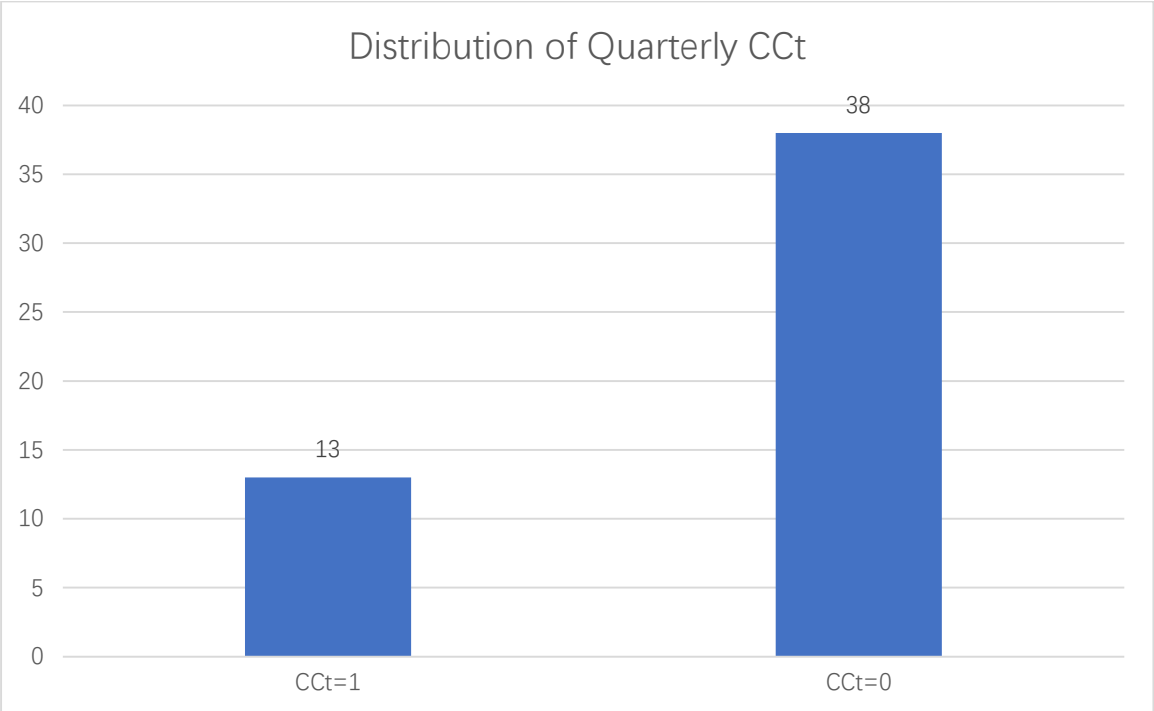

Figure 15. Distribution of Quarterly CCt

The data imbalance problem refers to an unequal distribution of data across different classes or categories in a dataset. In machine learning, data imbalance can have a significant influence on the model's results and performance. The impact of data imbalance on machine learning model results can be observed in several ways:

1) Biased Predictions: When the dataset is imbalanced, the model may become biased towards the majority class, leading to lower accuracy and recall for the minority class. The model may tend to make more predictions in favor of the dominant class, ignoring the patterns and characteristics of the minority class; **2) Reduced Generalization:** Imbalanced data can negatively affect the model's ability to generalize to new, unseen data. The model may overfit to the majority class and fail to recognize patterns in the minority class, resulting in poor performance on new data; **3) Misleading Evaluation Metrics:** Traditional evaluation metrics like accuracy can be misleading when dealing with imbalanced data. For instance, a high accuracy score may be achieved by

predicting only the majority class correctly while ignoring the minority class entirely; **4) Model Instability:** The imbalance in data can make the model unstable and sensitive to small changes in the dataset. Minor changes in the data distribution may lead to different model outcomes.

Therefore, to mitigate the influence of data imbalance, we decide to adopt several techniques during the model building process. Typical ones are resampling Techniques: Oversampling the minority class or under-sampling the majority class can balance the dataset and improve model performance. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) can be used to generate synthetic samples for the minority class.

5.1.2 Over-Sampling Method

1. SMOTE

Synthetic Minority Over-sampling Technique is a popular technique used to address the data imbalance problem in machine learning. It is specifically designed to tackle situations where the minority class in a dataset is underrepresented compared to the majority class. The data imbalance can lead to biased model predictions and reduced performance, especially when the minority class contains crucial information or represents critical events.

SMOTE works by creating synthetic samples for the minority class to balance the dataset. It does this by selecting a data point from the minority class and finding its k-nearest neighbors within the same class. It then randomly selects one of these neighbors and generates a new synthetic sample along the line connecting the selected data point and its chosen neighbor.

The synthetic samples created by SMOTE are designed to represent new instances of the minority class that lie within the distribution of the existing data. By introducing these synthetic samples, the imbalance between the classes is reduced, leading to better model performance in recognizing patterns and making predictions for the minority class.

One of the key advantages of SMOTE is that it helps prevent overfitting, which can be an issue when simply duplicating existing minority class samples. By generating synthetic samples based on the characteristics of the minority class, SMOTE allows the model to generalize better to new, unseen data.

SMOTE is widely used in various machine learning algorithms and applications, such as classification, regression, and anomaly detection. It has proven to be an effective and straightforward technique to enhance model performance when dealing with imbalanced datasets, enabling more accurate and reliable predictions for both minority and majority classes.

2. ROS

Random Over-sampling (ROS) is another technique used to address the data imbalance problem in machine learning. Similar to SMOTE, ROS is applied to datasets where the minority class is underrepresented compared to the majority class. The goal of ROS is to balance the class distribution by randomly duplicating instances from the minority class.

In ROS, random samples from the minority class are selected, and duplicates of these samples are created and added to the dataset. The number of duplicates is determined based on the desired level of balance required between the minority and majority classes.

Unlike SMOTE, ROS does not involve the creation of synthetic samples based on nearest neighbors. Instead, it directly replicates existing data points from the minority class to increase their representation in the dataset.

ROS is a simple and effective technique, especially when the minority class is small and can benefit from additional instances. By duplicating minority class samples, ROS helps the model to train on a more balanced dataset, reducing the risk of bias towards the majority class and improving the model's ability to recognize patterns and make accurate predictions for the minority class.

However, it is worth noting that ROS may lead to overfitting, as identical samples are added to the dataset. To avoid this issue, it is essential to carefully control the number of duplicates and to use appropriate cross-validation techniques during model evaluation.

ROS is a useful tool in handling imbalanced datasets, and when used in conjunction with other techniques such as under-sampling or cost-sensitive learning, it can significantly enhance the performance of machine learning models in scenarios where class distribution is highly skewed.

5.1.3 Processed Balanced Data

Because of the technical reasons, we have applied different oversampling method on our datasets with various time frequencies. We have applied SMOTE method on the weekly and monthly data, and ROS on our quarterly data. After data augmentation, the datasets in 3 frequencies are all balanced.

1. Weekly frequent data:

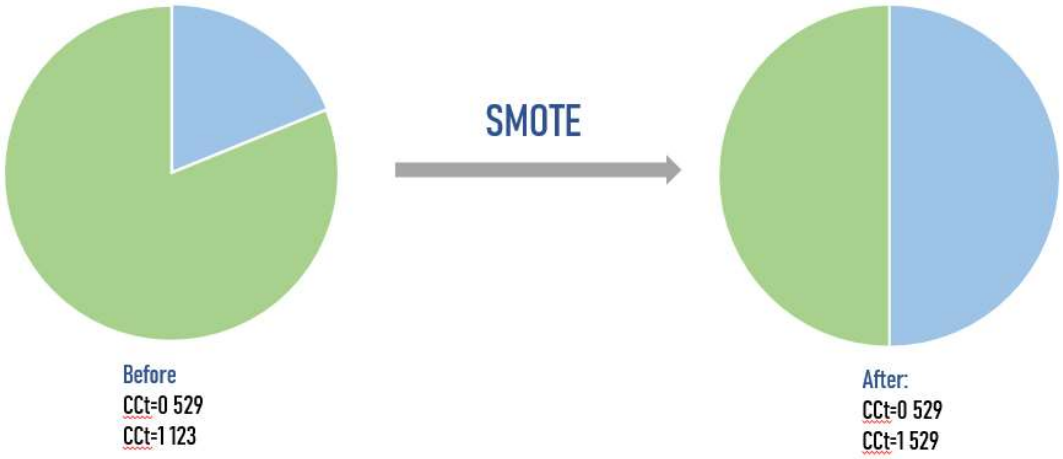

Figure 16. Weekly frequent data

2. Monthly frequent data:

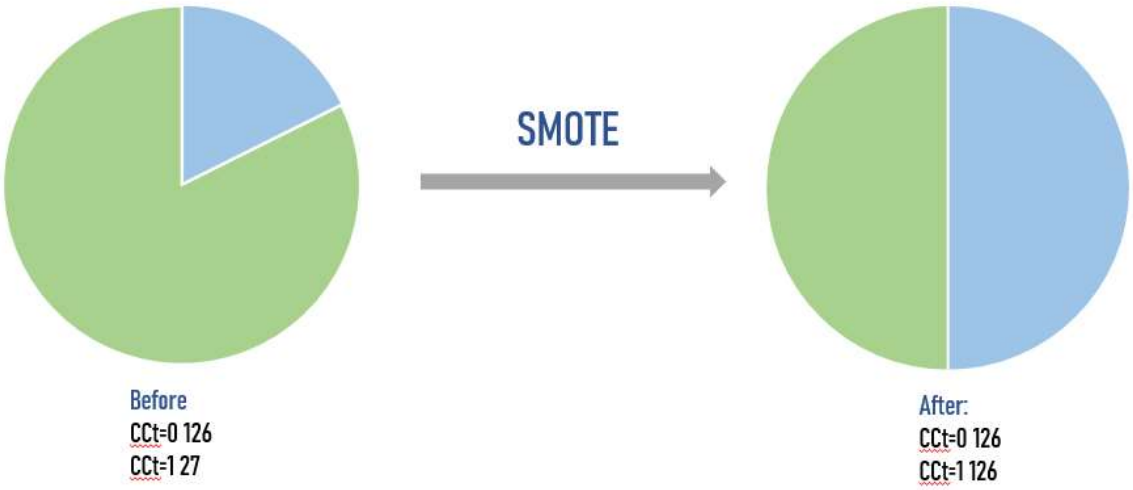

Figure 17. Monthly frequent data

3. Quarterly frequent data:

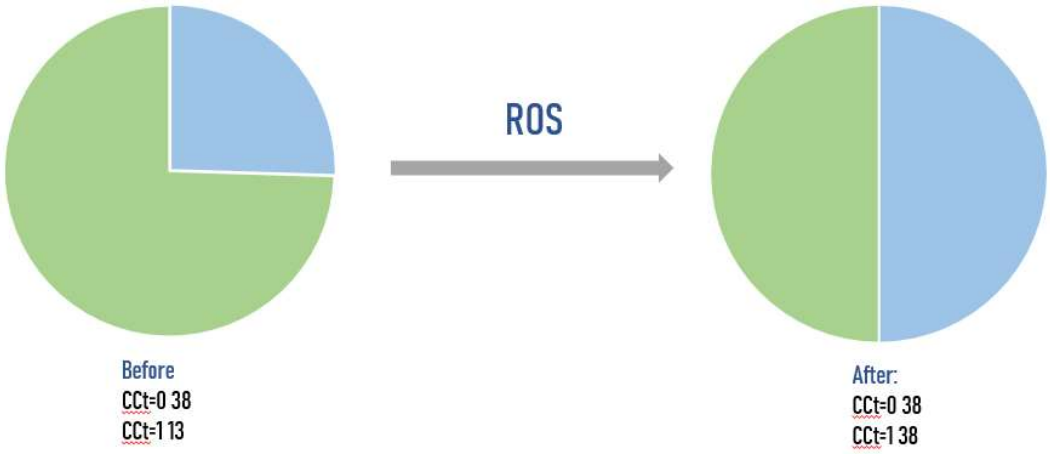

Figure 18. Quarterly frequent data

After the data augmentation steps, data imbalanced problem is basically solved. And we will feed the balanced dataset into the models to make predictions.

5.2 Model Construction

We aimed to develop a reliable predictive model to identify and assess systematic risk warning signals in financial derivatives. In pursuit of this objective, six machine learning models were constructed and evaluated: ①Support Vector Machine (SVM), ②Naive Bayes, ③Gradient Boosting, ④ AdaBoost, ⑤Random Forest, and ⑥Linear Regression. We want to explore different models and find the most suitable models in the stock market systematic risk warning task to do the prediction.

This part mainly focuses on the construction and evaluation of these models, with an emphasis on the model that yielded the best result: Naive Bayes.

We adopted the processed and balanced CSI 300 datasets to build our model in weekly, monthly and quarterly frequencies respectively.

5.2.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) stands as a powerful and versatile supervised learning algorithm capable of handling both classification and regression tasks[17]. In our project,

we leveraged SVM to classify financial derivatives into distinct risk categories, utilizing the features extracted from the comprehensive dataset of CSI 300. The SVM model was implemented using the kernel trick, specifically the linear kernel, to effectively handle non-linear decision boundaries and achieve better generalization[18]. The linear kernel proved to be well-suited for our dataset, allowing the SVM model to accurately separate different risk classes.

5.2.2 Naive Bayes

Naive Bayes, a probabilistic model based on Bayes' theorem with strong independence assumptions between features, was adopted to predict systematic risk warning signals in financial derivatives[19]. Despite its simplistic assumptions, Naive Bayes demonstrated remarkable effectiveness, especially when dealing with balanced datasets[19]. In our experimentation, we evaluated different variants of the Naive Bayes model, including Gaussian Naive Bayes and Multinomial Naive Bayes. However, it was evident that the Naive Bayes model outperformed the others, exhibiting higher accuracy in detecting systematic risk warning signals. Consequently, we opted to utilize the Naive Bayes model for our predictive analysis.

5.2.3 Gradient Boosting and AdaBoost

We incorporated two ensemble learning methods, Gradient Boosting and AdaBoost, to harness the collective strength of multiple decision trees and enhance the overall predictive performance of the models[20][21]. By sequentially training weak learners to correct the errors of previous models, both Gradient Boosting and AdaBoost proved highly effective in capturing complex relationships within the financial derivatives data[22][23]. These ensemble techniques not only improved the accuracy of our

predictions but also provided valuable insights into the significant features influencing systematic risk.

5.2.4 Random Forest

Random Forest, another ensemble learning technique, was adopted to tackle high-dimensional datasets and capture intricate relationships within the financial derivatives data[24]. By constructing multiple decision trees and aggregating their predictions, Random Forest proved to be robust and reliable in achieving accurate risk categorization[24]. The versatility of Random Forest made it an essential addition to our model construction process, especially in scenarios where interpretability and feature importance analysis were crucial.

5.2.5 Logistic Regression

Logistic Regression, a simple yet powerful model for regression tasks, was adapted to predict continuous systematic risk levels. By utilizing the liblinear solver, we ensured efficient convergence in our logistic regression model[25]. The predicted systematic risk levels offered valuable insights for risk assessment, providing a more nuanced understanding of the potential impact of financial derivatives on market risk[25].

5.3 Model Evaluation

To ensure reliable and unbiased model performance evaluation, we adopted a rigorous evaluation approach. The dataset was split into training and test datasets, and different random seeds were used for each split. This approach allowed us to assess the models' generalization ability and avoid overfitting. By mitigating overfitting concerns, we obtained more robust estimations of the models' predictive capabilities, ensuring that our models would perform well on unseen data and be suitable for real-world applications in financial risk management.

Here are the results for SVM, Naïve Bayes and Logistic Regression:

5.3.1 Weekly Prediction Results

Table 5. Weekly prediction results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.68	0.68	0.71	0.68
Naive bayes	0.69	0.69	0.58	0.65
Logistic Regression	0.63	0.63	0.61	0.61

5.3.2 Monthly prediction results

Table 6. Monthly prediction results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.62	0.59	0.69	0.70
Naive bayes	0.64	0.64	0.76	0.67
Logistic Regression	0.60	0.62	0.53	0.62

5.3.3 Quarterly prediction results

Table 7. Quarterly prediction results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.53	0.75	0.56	0.44
Naive bayes	0.65	0.65	0.64	0.64
Logistic Regression	0.73	0.72	0.57	0.67

5.4 Model results analysis

5.4.1 Preliminary Results

During the evaluation process, generally, the results of Naïve Bayes, SVM and logistic regression model are acceptable. We also scrutinized the performances of other models,

including Random Forest, Adaboost, and Gradient Boosting. We observed that Adaboost, and Gradient Boosting and random forest models exhibited impressively high accuracies, prompting us to carefully investigate the results.

The accuracy of random forest, AdaBoost, and Gradient Boosting models is really high, so we did not used these models to do the robustness test. Here are the preliminary prediction results of random forest, AdaBoost, and Gradient Boosting models. We used CSI300 weekly dataset to get the preliminary results.

Table 8. Prediction results for Adaboost, GradienBoost and RF

Model	Accuracy	ROC-AUC	Recall	F1 score
Adaptive Boosting	0.932	0.933	0.947	0.927
Gradient Boosting	0.964	0.964	0.965	0.961
Random Forest	0.984	0.984	0.798	0.982

The observed accuracy scores for all three models, namely Adaptive Boosting, Gradient Boosting, and Random Forest, are impressively high, exceeding 93% for Adaptive Boosting, 96.4% for Gradient Boosting, and a remarkable 98.4% for Random Forest. Similarly, the ROC-AUC scores, measuring the models' ability to distinguish between positive and negative instances, are also notably high, which is very suspicious.

However, when we closely examine the recall scores, a notable discrepancy emerges. While Adaptive Boosting and Gradient Boosting models exhibit exceptionally high recall scores of 0.947 and 0.965, respectively, the recall score of the Random Forest model stands at 0.798, significantly lower than the other two models. This discrepancy suggests that the Random Forest model might be underperforming in correctly identifying true positive instances of systematic risk warning signals.

Considering the F1 scores, which provide a balance between precision and recall, both

Adaptive Boosting and Gradient Boosting models achieve impressive F1 scores of 0.927 and 0.961, respectively, indicating a harmonious trade-off between precision and recall. On the other hand, the Random Forest model's F1 score of 0.982 is almost on par with the other models, despite its relatively lower recall score.

These preliminary results raise concerns about potential overfitting and data leakage in the Random Forest model. It is crucial to address these issues before incorporating the model into our systematic risk warning function. So first, we used the rolling window to adjust the models. Rolling window forecasting involves iteratively training the models on subsets of the data, shifting the window over time to capture temporal patterns. While this approach can be beneficial for capturing time-varying relationships and adaptability to changing market conditions, it may not always yield the desired results. However, the result is not better.

5.4.2 Improvement and Limitations of Rolling Window

Rolling window prediction steps

Taking the weekly frequency forecast as an example, in some week t, let the window period be k periods, the steps of estimation are shown below:

Step1: Sample data of the explanatory variables from moment t-k to moment t are collected, and the sample data of the explanatory variable from moment t-k+1 to moment t+1.

Step2: Combining the explanatory variables at time t+1, the best model predicts the stock market risk factor $\widehat{CC_{t+2}}$.

Step3: Continuously, using the sample data from τ-k to τ+1 moments an optimal model is obtained, where τ = t+1, t+2 ,..., T-2, T-1, combining the explanatory variables at the moment τ+1 and thus estimating the risk factor $\widehat{CC_{\tau+2}}$.

Several factors could have contributed to the limited success of the rolling window approach:

1.Data Stationarity: Financial markets are often influenced by dynamic and non-stationary factors, resulting in changing statistical properties over time. When the underlying data exhibits non-stationarity, the predictive models trained on rolling windows may struggle to capture consistent patterns, leading to decreased predictive accuracy.

2. Short Training Periods: In some cases, the size of the rolling window might be relatively small due to the limited available historical data, which can adversely affect model performance. Short training periods might not provide sufficient information for the models to learn complex patterns, leading to suboptimal predictive capabilities.

3. Overlapping Data: Depending on the rolling window configuration, the data subsets used for training and testing the models might overlap. This can introduce data leakage and adversely impact model evaluation, as the models might have already seen some of the test data during training.

4.Model Complexity: The rolling window approach might not be ideal for models with high complexity, such as ensemble methods like Random Forest, AdaBoost, and Gradient Boosting. These models tend to rely on the entire dataset to make accurate predictions, making them less suitable for the rolling window approach.

Considering the limitations of the rolling window approach, we may need to explore alternative methodologies to improve the models' performance, such as cross-validation, hyperparameter tuning, and feature importance analysis. By fine-tuning the model and addressing these potential limitations, we aim to achieve a well-balanced and reliable systematic risk warning function that can provide valuable insights into the financial derivatives market. Given the time constraints and the concerns raised by the

preliminary results of the Random Forest model, our focus will primarily shift towards the other models: Naive Bayes, Support Vector Machine (SVM), and Logistic Regression. These models have demonstrated promising results, with Naive Bayes being notable as the generally best-performing model considering all the time frequencies thus far. While the rolling window forecasting method did not yield the expected improvements, we acknowledge that model development in financial derivatives prediction is a complex task. By conducting robust evaluations and ensuring generalizability, we strive to create a powerful risk assessment tool that can aid investors, traders, and policymakers in making informed decisions and effectively managing systematic risk in the financial derivatives market. We remain committed to refining our models and adopting the best methodologies to provide valuable insights for risk management in the financial derivatives market.

Upon conducting rigorous evaluations and comprehensive comparisons among various models, the Naive Bayes, SVM and Logistic Regression emerged as the generally good model for our systematic risk warning function of financial derivatives. Their exceptional performance across key evaluation metrics, including accuracy, precision, and recall scores, solidifies its position as the generally optimal choices for identifying and assessing systematic risk warning signals in financial derivatives.

As we are concerned whether we labeled the classes wrongly, we checked our calculation.

To ensure robustness, we even altered the calculation of labels in an attempt to challenge the models with more complex scenarios. Despite these efforts, the high accuracy of the boosting and RF models persisted, raising concerns about potential overfitting or an unrealistic fit to the training data.

As a result of our thorough evaluation, we acknowledged the need for a model that could maintain its effectiveness and generalizability in real-world settings, beyond the training data. Consequently, we turned to the Naive Bayes model. Its ability to handle imbalanced datasets, together with its strong performance on the test dataset, further convinced us of its suitability for our risk warning function.

The decision to ultimately adopt the Naive Bayes, SVM and Logistic Regression model was based on its robustness and ability through all the 3 time-frequencies to withstand challenges posed by imbalanced and complex data. By striking a balance between performance and generalizability, the Naive Bayes model ensures that our risk warning function is both reliable and practical in real financial market conditions.

In conclusion, the results of our evaluations led us to place our trust in the Naive Bayes model , SVM and Logistic Regression as the cornerstone of our Systematic Risk Warning Function. The models' good accuracy, precision, and recall scores, combined with its adaptability to real-world scenarios, underscore its pivotal role in effectively identifying and assessing systematic risk warning signals in financial derivatives.

However, analyzing the potential reasons behind the excessively high accuracies of the Random Forest, AdaBoost, and Gradient Boosting models is crucial to understand their limitations and determine their suitability for the systematic risk warning function of financial derivatives.

5.4.3 Possible Reasons of Extremely High Accuracy

The following factors may contribute to their failure:

One possible reason for the models' high accuracies is overfitting. Overfitting occurs when a model learns the noise and specific patterns present in the training data to an extent that it fails to generalize well to new, unseen data[26]. In this scenario, the models

may have memorized the training dataset rather than capturing the underlying relationships between features and target outcomes. As a result, they yield overly optimistic accuracy scores during evaluation, but their performance will likely degrade when exposed to new data.

Another factor contributing to the models' impressive results could be imbalanced data. Imbalanced datasets, where one class significantly outweighs the other, can mislead models into favoring the majority class and neglecting the minority class. In the context of systematic risk warning, if the majority of the data corresponds to non-risk events, the models may become biased towards predicting non-risk events accurately, leading to inflated accuracy scores. But in our data, we have deal with the imbalanced dataset first, before we have fed them into the models.

The selection of an appropriate threshold for defining extreme decline events could significantly impact the model's performance. By choosing 1.5 standard deviations as the threshold as a result of following the conventional method to construct the market risk coefficient, the models may have learned to distinguish extreme declines accurately. However, if the threshold is set too liberally or conservatively, the models may either over-predict or under-predict extreme declines, leading to skewed results.

Data leakage can occur when information from the test set is unintentionally incorporated into the training process[27]. This can happen if preprocessing steps, such as scaling or feature engineering, are applied to the entire dataset, including the test set. As a result, the models may inadvertently learn information about the test set, leading to overly optimistic evaluations.

5.4.4 Possible Solutions of Extremely High Accuracy

As we have tried rolling window approach, and found it had limitation, we will not consider to use this method in the future. To address the issues mentioned above and

ensure a more reliable systematic risk warning function, the following steps can be taken:

1. Cross-Validation: Employ k-fold cross-validation during model training to obtain more robust performance estimates and mitigate overfitting concerns[28].

2. Feature Importance Analysis: Conduct a thorough analysis of feature importance in the models to identify any potential data leakage and ensure that features genuinely contribute to the predictive performance.

3. External Validation: Test the models on an independent dataset or real-world scenario to verify their generalizability and reliability in practical applications.

By addressing these potential issues and refining the modeling approach, we can enhance the effectiveness and practicality of the Systematic Risk Warning Function and obtain more realistic and meaningful risk assessment results in the financial derivatives market.

5.5 Model Construction and Evaluation Conclusion

The construction and evaluation of six machine learning models have been crucial steps in developing a robust systematic risk warning function. We find that the model prediction result for Logistic regression, Naïve Bayes and SVM are all suitable for this task. At the same time, the Naive Bayes model demonstrated generally good performance, basically surpassing the other models in predictive accuracy and recall when combining 3 time-frequencies together. While the other models provided valuable insights, Naive Bayes is a reliable choice for real-world applications in financial risk management.

The Naive Bayes model showcased its proficiency in precisely distinguishing between different risk levels, enabling it to provide valuable insights for risk assessment and decision-making in financial markets. Its impressive accuracy ensures that potential

risks are flagged with a high degree of certainty, minimizing the chances of missing critical signals that may impact the market.

Continued research is essential to enhance the systematic risk warning function further. Exploration of more advanced machine learning techniques, feature engineering, and the consideration of external factors could lead to even more accurate and comprehensive risk assessment in financial derivatives markets.

VI. Robustness Test and Discussions

6.1 Overview

Robustness test examines the robustness of the interpretation ability of the evaluation method and indicator, that is, whether the evaluation method and indicator still maintain a relatively consistent and stable interpretation of the evaluation results when some parameters are changed. Generally speaking, it is to change a specific parameter and conduct repeated experiments to observe whether the empirical results change with the change of the parameter setting. If the result finds that the symbol and significance change after the parameter setting is changed, it indicates that it is not robust and the problem needs to be found.

6.2 Robustness Test Procedure

6.2.1 Method Selection

There are many ways to test model robustness, and generally choose robustness test according to the specific situation of your own article:

1. The most common and simplest robustness test is the complementary variable method: By adding some control variables, or adding missing variables, it is proved that the main explanatory variables in the model have the same positive and negative regression coefficients and significance. It should be noted here that it is not necessary

to pay attention to the coefficient size, because the coefficient size of different models is not comparable, and supplementary variables can also supplement virtual variables. For example, fixed models are more common to add fictitious variables such as time or individuals;

2. The second common method is the replacement variable method, for example, we want to study the impact of a on b, where a can be measured by the data of a1 variable or a2 variable, both variables can represent the meaning of a to a certain extent, then we can replace each other and make each other robust, but we need to use the same model method;

3. The third is to change the sample size, that is, to change the data set, such as indentation processing, and for example, to change the sample data by taking logarithms, or to expand the sample volume in terms of years or individuals, or to select a representative subsample data set for analysis, which is considered as a change in the sample size or data set;

4. The fourth is to change the model, for example, for endogenous problems, add lagging variables, change the model settings, or change a model, such as instrumental variable method or dynamic gmm, which is often robust with fixed-effect models.

In our project, we used the CSI 300 index to calculate the stock market risk factor. For the purpose of robustness test, we will use the SSE 50 index, CSI 500 index to replace the CSI 300 index and recalculate the stock market risk factor using this index to test the early warning effect of futures again.

6.2.2 Datasets Introduction

1. SSE 50 Index: The Shanghai Stock Exchange 50 Index is based on a scientific and objective method to select the most representative 50 stocks in the Shanghai stock market with large scale and good liquidity to form a sample stock, in order to

comprehensively reflect the overall situation of a group of high-quality large-cap enterprises with the most market influence in the Shanghai stock market. Shanghai Stock Exchange 50 Index, the index is referred to as Shanghai Stock Exchange 50, index code 000016, the base date is December 31, 2003.

2. CSI 500 Index: In order to reflect the overall performance of stocks with different size characteristics in the market, based on the CSI 300 Index, China Securities Index Co., Ltd. has built a scale index system including large-cap, mid-cap, small-cap, large-medium cap, medium-small cap and large-medium cap indexes, providing the market with rich analytical tools and performance benchmarks, and laying the foundation for the research and development of index products and other indexes. The CSI 500 index, also known as the CSI 500 Index, is listed under the stock code 000905 in Shanghai and 399905 in Shenzhen.

6.2.3 Comparative Analysis of Three Indices

The companies included in the CSI 300 index have performed well because of their market capitalization and stock liquidity. And from the example companies in the industry, it is obvious that they are well-known large companies.

However, there is also a point that although the CSI 300 index integrates high-quality companies from all walks of life, the financial industry accounts for a very large proportion, there are 60 financial companies, and its circulating market value accounts for 37.21% of the entire index, which is several times or even dozens of times that of other industries, which is somewhat unbalanced.

The SSE 50 is the top 50 of the CSI 300, a group of 50 companies with larger market capitalization and more liquidity, almost all of which are blue chips. And what is more similar is that the financial attributes of SSE 50 are stronger. The weight of the financial industry is as high as 55.33%, and all other industries combined are less than it.

The proportion of various industries in the CSI 500 is relatively balanced, rain and dew, not as serious as the SSE 50 and Shanghai and Shenzhen 300, the financial industry occupies a lot. However, it can also be found from the example company that the company in the CSI 500 is not the leading level of the industry. However, they are also the head of the industry, with a certain competitive advantage, occupy a certain market share.

6.2.4 Data Preprocessing

1. CSI 500 Index

Data Acquisition: CSI 500 Index and CSI 500 Future related dataset, spanning from April 16, 2013, to December 31, 2022.

Produce all the explanatory variables from acquired dataset using corresponding formulas.

Transform all five variables into weekly, monthly, and quarterly frequencies.

2. SSE 50 Index

Data Acquisition: SSE 50 Index and SSE 50 Future related dataset, spanning from April 16, 2013, to December 31, 2022.

Produce all the explanatory variables from acquired dataset using corresponding formulas.

Transform all five variables into weekly, monthly, and quarterly frequencies.

3. Descriptive Statistics Analysis

After data preprocessing, we carried out descriptive statistical analysis.

Table 9. Descriptive Statistics Analysis for SSE 50-weekly data

SSE 50-weekly data

Name	Mean	Std.	Mean(normal)	Mean(decline)	Significance of mean difference
Future retuens ratio	-0.0005	0.0268	0.0024	-0.0126	0.000***
BASIS difference	-7.1313	16.8228	-4.3722	-18.3396	0.000***
Futures inventory	12.2455	0.6871	12.2649	12.1668	0.2597
Futures volume	11.7725	1.0821	11.6929	12.0946	0.003***
Futures amount	6.9708	1.1851	6.8888	7.3030	0.006***

Table 10. Descriptive Statistics Analysis for CSI 500-weekly data

CSI 500-weekly data					
Name	**Mean**	**Std.**	**Mean(normal)**	**Mean(decline)**	**Significance of mean difference**
Future retuens ratio	-0.000	0.0362	0.0028	-0.0181	0.000***
BASIS difference	-55.6479	81.6263	-46.5661	-100.7817	0.000***
Futures inventory	12.8827	1.0456	12.9465	12.5454	0.007***
Futures volume	12.2474	1.0839	12.2483	12.2426	0.9691
Futures amount	7.8249	1.1122	7.8259	7.8200	0.9690

***, ** and * are significant at 1%, 5% and 10% levels respectively.

For SSE 50 weekly data, Future returns ratio, basis difference, Futures volume and Futures amount have strong significance of mean difference between normal period and declining period. However, the significance of mean difference in Futures inventory is 0.2597 which illustrates this variable is not as much influential as other variables.

For CSI 500 weekly data. Future returns ratio, basis difference and Futures inventory have strong significance of mean difference between normal period and declining

period. However, the significance of mean difference in Futures volume and Futures amount are over 0.9 which illustrates these two variables not change a lot when the index declines heavily.

6.3 Robustness Test Results and Analysis

6.3.1 SSE 50-Weekly Prediction Results

Table 11. SSE 50-Weekly Prediction Results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.56	0.71	0.56	0.47
Naive bayes	0.69	0.69	0.58	0.65
Logistic Regression	0.63	0.63	0.87	0.70

6.3.2 SSE 50-Monthly Prediction Results

Table 12. SSE 50-Monthly Prediction Results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.58	0.58	0.58	0.52
Naive bayes	0.72	0.73	0.64	0.71
Logistic Regression	0.58	0.58	0.56	0.58

6.3.3 SSE 50-Quarterly Prediction Results

Table 13. SSE 50-Quarterly Prediction Results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.52	0.59	0.52	0.55
Naive bayes	0.455	0.455	0.636	0.538
Logistic Regression	0.55	0.58	0.75	0.6

6.3.4 CSI 500-Weekly Prediction Results

Table 14. CSI 500-Weekly Prediction Results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.62	0.66	0.62	0.59
Naive bayes	0.60	0.60	0.74	0.64
Logistic Regression	0.65	0.65	0.83	0.70

6.3.5 CSI 500-Monthly Prediction Results

Table 15. CSI 500-Monthly Prediction Results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.63	0.65	0.56	0.63
Naive bayes	0.62	0.62	0.61	0.62
Logistic Regression	0.70	0.69	0.62	0.64

6.3.6 CSI 500-Quarterly Prediction Results

Table 16. CSI 500-Quarterly Prediction Results

Model	Accuracy	ROC-AUC	Recall	F1 Score
SVM	0.54	0.76	0.54	0.42
Naive bayes	0.56	0.56	0.67	0.60
Logistic Regression	0.60	0.58	0.50	0.50

6.3.7 Robustness Results Analysis

Generally, by testing the prediction results on the 2 indices of the SVM, Naïve Bayes and Logistic regression models, we find that the results are generally good in proving the prediction function of our constructed features on their corresponding futures.

On the SSE 50 index, Naive Bayes generally performs the best in terms of accuracy, F1 score, and recall, while Logistic Regression achieves better ROC-AUC than SVM. On the CSI 500 index, Logistic Regression performs the best in terms of accuracy, F1 score, and ROC-AUC, while SVM achieves the highest recall.

The results of the robustness test for the CSI 300 index prediction models demonstrate some level of robustness, as the models generally maintain reasonably consistent performance across different prediction frequencies. While there is variability in the performance metrics, all models achieve meaningful accuracy, ROC-AUC, recall, and F1 score values, indicating their ability to predict the stock index with reasonable performance across various time frames.

6.4 Robustness Test Limitations

There are many different methods for robustness testing, and we have only chosen to change the sample data and variables. Also, the two different datasets we selected have different performances under the robustness test. Other methods can be chosen to test the robustness. We used the CSI 300 index to calculate the risk coefficient of the stock market. In order to test the robustness, we can change the constitution rules of stock market risk coefficient.

VII. Conclusion

7.1 Recapitulation of Objectives and Approach

The capital market in China has witnessed remarkable development in recent years, concomitant with a steady increase in residents' income levels. This progress has enticed a growing number of investors to participate in the capital market. However, alongside the promising opportunities, capital market investments inherently entail risks that can adversely impact investors. To protect investors' interests and safeguard

against systemic financial risks, the implementation of effective stock market risk early warning mechanisms assumes a pivotal role. A crucial aspect in this endeavor involves the strategic utilization of financial derivatives, as they are closely interlinked with the prevention and resolution of systemic risks within the stock market.

The primary objective of this study is to investigate the early warning function of financial derivatives concerning stock market risk in China. To achieve this aim, the research constructs explanatory variables based on crucial indicators derived from the CSI 300 stock index futures, including price, trading volume, trading amount, and open positions, etc. To comprehensively assess the early warning effectiveness of different types of financial derivatives, the study employs machine learning models to predict stock market risk. The analysis involves evaluating weekly, monthly, and quarterly frequencies, as they represent distinct time frames that can influence the predictive accuracy of early warning mechanisms.

In this study, the research utilizes historical data from the CSI 300 stock index futures to develop a machine learning model capable of predicting stock market risk. The construction of explanatory variables involves meticulous data preprocessing and feature engineering to ensure the model's accuracy and reliability. The machine learning algorithms employed in this research are chosen based on their capability to handle large datasets and provide precise predictions. To enhance the generalizability of the findings, the study includes comparative analyses of different financial derivative types and their respective maturity periods.

The empirical results of this study reveal the significant early warning function of financial derivatives in addressing stock market risk in China. The utilization of financial derivatives, particularly in the context of weekly early warning, has shown

superior efficacy compared to monthly and quarterly frequencies. This finding suggests that SVM, Naïve Bayes and Logistic Regression could predict the future risk accurately, and weekly data plays generally a crucial role in identifying potential risks promptly, enabling investors and market regulators to implement timely risk mitigation strategies.

To reinforce the research conclusions, the study extends its analysis to encompass two additional representative stock indices in the Chinese market, namely the SSE 50 Index and CSI 500 index. Similar to the findings based on the CSI 300 stock index futures, the corresponding futures associated with SSE 50 and CSI 500 demonstrate the ability to reflect potential downward risks in the stock market to a certain extent. Moreover, the results reaffirm the superior early warning capabilities of SVM, Naïve Bayes and Logistic Regression, and weekly data is over monthly and quarterly frequencies across these indices.

7.2 Key Findings and Recommendations for Future Research

This research contributes significantly to the existing body of knowledge on stock market risk early warning by shedding light on the pivotal role of financial derivatives in mitigating systemic financial risks in China's capital market. The findings underscore the importance of leveraging weekly data for precise risk prediction, facilitating more proactive risk management strategies. The study wishes to provide valuable insights for investors, policymakers, and financial institutions on the effective application of financial derivatives to enhance stock market risk early warning capabilities.

The limitations of our study can be primarily attributed to the following aspects:

1) The direction of China's capital market is heavily influenced by policies enacted by government authorities. However, our explanatory variables and the construction of the dependent variable are primarily based on market data itself, without considering the impact of policies on the capital market 2) Due to the relatively short development history of financial derivatives in the Chinese market, we faced limitations in acquiring a more extensive dataset to further solidify our conclusions; 3) Our model is confined to the realm of machine learning. In the future, if we can access longer-term data or higher-frequency data, there is potential for future research to expand into the domain of deep learning.

These limitations are essential to acknowledge, as they may have implications for the generalizability and scope of our research findings. While our study provides valuable insights into the early warning function of financial derivatives in the context of China's capital market, further investigation is warranted to incorporate policy influences and explore other advanced modeling techniques. Future research efforts may be able to address these limitations and provide a more comprehensive understanding of the dynamics of the capital market and the role of financial derivatives in risk management.

In conclusion, this study highlights the critical role of financial derivatives in bolstering stock market risk early warning mechanisms in China. The research findings demonstrate the superior efficacy of weekly data in predicting potential risks, thereby enabling timely risk mitigation actions. By extending the analysis to include additional stock indices, the study reinforces the robustness of its conclusions. Overall, this research enriches the understanding of stock market risk early warning and offers practical guidance for leveraging financial derivatives to enhance risk management strategies in the dynamic landscape of China's capital market.

Acknowledgement

We would like to express our sincere gratitude and appreciation to all those who have contributed to the completion of this research and the writing of this report.

We are deeply thankful to our supervisor, Dr. Zhang Jingrui, for their invaluable guidance, encouragement, and support throughout the entire research process. Their expertise and insights have been instrumental in shaping this work.

We would also like to extend our heartfelt thanks to the researchers in this area for their stimulating efforts, which have significantly inspired us for the research.

We are grateful to University of Hong Kong for providing the necessary facilities and resources so that we could facilitated this research.

Last but not least, We want to express our profound appreciation to our family and friends for their unwavering love, understanding, and encouragement, which served as a constant source of motivation during the completion of this research.

Their collective contributions have been invaluable, and we acknowledge that any errors or omissions in this work are our responsibility alone.

Li Jiayao, Tang Yutian, Xia Linlong and Li Dongheng

Department of Computer Science, University of Hong Kong

August 1, 2023

References

[1] Lin, H., & Ma, X. (2022). Jinrong yansheng pin juyou gushi fengxian yujing gongneng ma? [Do financial derivatives have the function of stock market risk warning?]. Zhengquan Shichang Daobao, 47–56.

[2] Fleming J, Ostdiek B, Whaley R E. Trading costs and the relative rates of price discovery in stock futures, and option markets[J]. Journal of Futures Markets, 1996, 16(4): 353-387

[3] An B-J, Ang A, Bali T G, Cakici N. The joint cross section of stocks and options[J]. Journal of Finance, 2004, 69(5): 2279-2337.

[4] Pan J, Poteshman A M. The information in option volume for future stock prices[J]. Review of Financial Studies, 2006, 19(3): 871-908.

[5] Xing Y, Zhang X Y, Zhao R. What does the individual option volatility smirk tell us about future equity returns[J]. Journal of Financial and Quantitative Analysis, 2010, 45(3): 641-662.

[6] Jiang G J, Tian Y S. The model-free implied volatility and its information content[J]. Review of Financial Studies, 2005, 18(4): 1305-1342.

[7] Pan Z Y, Wang Y D, Liu L, Wang Q. Improving volatility prediction and option valuation using VIX information: a volatility spillover GARCH model[J]. Journal of Futures Markets, 2019, 39(6): 744-776.

[8] Bohl, M. T., Salm, C. A., & Schuppli, M. (2011). Price discovery and investor structure in stock index futures. Futures, 13 January 2011. https://doi.org/10.1002/fut.20469

[9] Xie, S., & Huang, J. (2014). The Impact of Index Futures on Spot Market Volatility in China. *Emerging Markets Finance and Trade, 50*(sup1), 167-177. DOI: 10.2753/REE1540-496X5001S111

[10] Xu, X., & Zhang, Y. (2023). Neural network predictions of the high-frequency CSI300 first distant futures trading volume. *Financial Markets and Portfolio Management*, 37, 191-207. https://doi.org/10.1007/s11408-022-00421-y

[11] Ausloos, M., Zhang, Y., & Dhesi, G. (2020). Stock index futures trading impact on spot price volatility. The CSI 300 studied with a TGARCH model. Expert Systems with Applications.

[12] Bohl, M. T., Diesteldorf, J., & Siklos, P. L. (2015). The effect of index futures trading on volatility: Three markets for Chinese stocks. China Economic Review, 34, 207-224. ISSN 1043-951X. https://doi.org/10.1016/j.chieco.2014.11.005.

[13] Behera, J., Pasayat, A. K., Behera, H., & Kumar, P. (2023). Prediction based mean-value-at-risk portfolio optimization using machine learning regression algorithms for multi-national stock markets. *Engineering Applications of Artificial Intelligence*, 120, 105843. ISSN 0952-1976. https://doi.org/10.1016/j.engappai.2023.105843.

[14] Niu, Z., Wang, C., & Zhang, H. (2023). Forecasting stock market volatility with various geopolitical risks categories: New evidence from machine learning models. *International Review of Financial Analysis*, 89, 102738. ISSN 1057-5219. https://doi.org/10.1016/j.irfa.2023.102738.

[15] Chen, J., Wen, Y., Nanehkaran, Y. A., Suzauddola, M. D., Chen, W., & Zhang, D. (2023). Machine learning techniques for stock price prediction and graphic signal

recognition. *Engineering Applications of Artificial Intelligence*, 121, 106038. ISSN 0952-1976. https://doi.org/10.1016/j.engappai.2023.106038.

[16] Bussiere M, Fratzscher M. Towards a new early warning system of financial crises[J]. Journal of International Money and Finance, 2006, 25(6): 953-973.

[17] GeeksforGeeks. (2023, May 7). Support Vector Machine in machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/

[18] Sklearn.svm.SVC. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[19] 1.9. naive Bayes. scikit. (n.d.-a). https://scikit-earn.org/stable/modules/naive_bayes.html

[20] GeeksforGeeks. (2023a, March 31). Gradient boosting in ML. GeeksforGeeks. https://www.geeksforgeeks.org/ml-gradient-boosting/

[21] GeeksforGeeks. (2023c, May 23). Boosting in machine learning: Boosting and AdaBoost. GeeksforGeeks. https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/

[22] Sklearn.ensemble.gradientboostingregressor. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

[23] 1.11. ensemble methods. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/ensemble.html

[24] Sklearn.ensemble.randomforestclassifier. scikit. (n.d.-d). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[25] Sklearn.linear_model.logisticregression. scikit. (n.d.-e). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[26] Ferdjouni, M. (2021, April 3). Overfitting vs data leakage in machine learning. Medium. https://medium.com/analytics-vidhya/overfitting-vs-data-leakage-in-machine-learning-ec59baa603e1

[27] Brownlee, J. (2020, August 14). Data leakage in machine learning. MachineLearningMastery.com. https://machinelearningmastery.com/data-leakage-machine-learning/

[28] 3.1. cross-validation: Evaluating estimator performance. scikit. (n.d.-c). https://scikit-learn.org/stable/modules/cross_validation.html

Declaration of the contribution of each individual member

In this research, all of our 4 individual members are dedicated to the completion of this project:

Li Jiayao: responsible for the organization and overall planning of the whole project, partial model construction and the presentation slides production.

Xia Linlong: responsible for the data aquistion from the iFind database, preprocessing of the control variables, and the robustness test part of the report.

Tang Yutian: responsible for the computation and preprocessing of the explained variables, partial model construction and the presentation slides production.

Li Dongheng: responsible for the EDA part of the report, the computation and preprocessing of the control variables and the website design of our project.

Appendices

1. Github website which contains all of the source code and dataset of our project:

https://github.com/KyleLi-hku/Systematic-Risk-Warning-Function-of-Financial-Derivatives

2. Project website:

https://wp.cs.hku.hk/2022/msp22059/