Final Year Report

# Market Manipulation Detection Using Supervised Learning

FITE4801 Final Year Project

Department of Computer Science

University of Hong Kong

| Group fyp23058 | UID |
|---|---|
| Ho Megan Qian Hua | 3035832749 |
| Tsang Hoi Wei | 3035785879 |
| Li Wo Him | 3035783986 |

Supervisor

Dr. Liu, Qi

Assistant Professor of Computer Science

21 January 2024

# ABSTRACT

The efficiency of financial markets rests upon the accurate reflection of supply and demand, bolstering sound investment decisions and fostering market stability. Despite regulatory measures, instances of market manipulation persist. This has posed a challenge for the Securities and Exchange Commission (SEC) in effectively detecting, analyzing, and documenting fraudulent activities. Gantz (2022) highlights the limitations to the reliance on predefined patterns for fraud detection, impeding the identification of emerging, unfamiliar patterns.

This paper introduces an approach employing supervised learning models to detect unidentified market manipulation patterns within publicly listed companies. First, data is collected from companies involved in stock manipulation from the CSRC's database. Subsequently, data cleaning and preprocessing are performed to prepare the data for comprehensive model analysis.

Four distinct machine learning models, including Support Vector Machine (SVM), Naïve Bayes, Decision Tree (DT) and Logistic Regression (LR), are built to evaluate their performances in identifying manipulation. Hyperparameters were repeatedly tuned to optimize each model's performance before comparing it with each other. The obtained results are expected to show that the SVM model outperforms the others, efficiently recognizing historically manipulated activities. However, while the SVM model showcases promising performance based on historical data, its real-time predictive capabilities are constrained by the absence of robust sentiment analysis on news and public information. Prospective research endeavors aim to augment predictive accuracy by integrating analyses of both insider and outsider information, envisaging an enhanced capacity to predict and preempt market manipulation.

# ACKNOWLEDGMENT

We extend our heartfelt gratitude to Dr. Liu Qi and Dr. Zou, Difan for their invaluable guidance and mentorship during this research project. We also acknowledge the financial support provided by the Department of Computer Science at The University of Hong Kong. Finally, we are deeply appreciative of the unwavering encouragement and support from our families and friends throughout our study.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# ABBREVIATIONS

| | |
|---|---|
| AUC | Area Under the Curve |
| CSRC | China Securities Regulatory Commission |
| DT | Decision Trees |
| LR | Logistic Regression |
| ROC | Receiver Operating Characteristics |
| SEC | Securities and Exchange Commission |
| SFC | Securities and Futures Commission |
| SMOTE | Synthetic Minority Oversampling Technique |
| SVM | Support Vector Machines |

## SECTION 1 – INTRODUCTION

The relentless pursuit of fast and seamless technological advancements has undeniably brought numerous benefits to humanity. However, as our world becomes increasingly sophisticated, it simultaneously offers more avenues for individuals with fraudulent intentions to exploit. Fraud has become an alarmingly pervasive issue, occurring at an astonishing rate of approximately every 15 seconds (Kottasova, 2016). This alarming statistic places individuals and businesses at constant risk of becoming victims of fraudulent activities, particularly within the finance industry. Market manipulation, an act of market abuse that disrupts the free and fair operation of markets, remains a significant concern for both regulators and investors. In Hong Kong, the Securities and Futures Ordinance (SFO) categorizes market manipulation as one of the offenses under market misconduct, subject to both civil and criminal regulations (Charltons, 2022). Market misconduct encompasses a wide range of activities, including false trading, price rigging, stock market manipulation, and the dissemination of false or misleading information to induce transactions (see Appendix A). Individuals involved in these actions, whether directly or indirectly, are subject to criminal prosecution (see Appendix B).

**The current approach to detecting market manipulation primarily involves analysing daily trading trends.** The focus of the current market surveillance system is on sudden fluctuations or unusual behaviours in share prices or trading volumes. Information from both internal and external sources is considered when evaluating the data (JPX, 2023). However, this investigation process still heavily relies on manual tracking and rule-based systems. The methods used may need to become more efficient in detecting manipulation. For example, JPMorgan Chase faced a fine of US$200 million for violating federal securities laws by allowing unapproved communications dating back to 2015. This has hindered the regulators from monitoring exchanges between banks and clients until 2021 (Franck & Son, 2021). Current reports have also proposed theoretical models for detecting market manipulation (Liu et al., 2021; Li et al., 2017), with machine learning models proving superior to traditional statistical approaches. Reflecting the evolving nature of market manipulation, Yi et al. (2023) developed a nonlinear model with a loss function as the training metric, achieving the highest accuracy in fraud detection within existing literature.

**This paper contributes in two ways.** First, while most scholars have focused on fine-tuning specific machine learning models, such as Support Vector Machines (SVM), there remains a need for more efficient and accurate methods for detecting market manipulation. To address this, we build and assess the performance of four machine learning models: Support Vector Machines (SVM), Decision Trees (DT), Naïve Bayes and Logistic Regression (LR). These models utilize supervised learning algorithms, with hyperparameters aligned with the latest regulations proposed by the Hong Kong Exchanges and Clearing Limited (HKEX) and SFC. Second, this paper will compare the performances to provide a comprehensive analysis of these models. We will fine-tune each model using additional hyperparameters and statistical techniques to optimize overall performances. The performance evaluation will be based on the receiver operating characteristic (ROC) and the area under the curve (AUC), with results presented on a website for enhanced visualization.

**The greatest challenge in detecting market manipulation lies in adapting to its evolving characteristics.** The introduction of new techniques and technologies has not only given rise to short-term market manipulation (Liu et al., 2021), but also the involvement of fraud in more than one market with a disguised identity. These challenges have further complicated the detection methods. To **tackle these challenges**, we incorporate machine learning models with techniques like the synthetic minority oversampling technique (SMOTE) to generate synthetic samples and introduce some noisy instances. SMOTE aims to balance imbalanced datasets, thereby improving the model's ability to detect manipulation within the system.

**The subsequent sections of this paper are organized as follows.** Section 2 introduces the methodology used for data and model development. Section 3 presents the fine-tuning of each machine learning model and the relevant experimental results, along with the proposed timeline of the project. Section 4 summarizes the work and outlines potential avenues for future research.

## SECTION 2 – METHODOLOGY

This chapter outlines the approach to dataset analysis and introduces the theoretical structures of the model.

### 2.1 Data Collection & Origination

The data used in this paper is the companies listed on the Shanghai Stock Exchange and the Shenzhen Stock Exchange, focusing on those with violations related to "Market Manipulation" or "Stock Price Manipulation" (See Table 1). To account for the typical three to four-year delay between the investigation and the announcement of market manipulation, the dataset covers a three-year period from January 2020 to December 2022. The selection of this time frame was deliberate, considering the latest revision of the security law in 2019. This strategic choice enables a more inclusive analysis, as a broader spectrum of stocks align with the revised laws, administrative regulations, and rules. The intention is to facilitate a comprehensive examination of the data, considering the regulatory landscape and legal framework in place during the specified period. The dataset by CSRC consists of a total of 2,600 unique companies being selected for market manipulation analysis. Among the raw data, there are 1508 negative samples and 1273 positive samples. It seems like a balanced dataset upon initial data collection. Violation may or may not be market manipulation. Given there is presence of imbalanced data, understanding the dataset is crucial as it will significantly impact the performance of the machine learning models. Insights and guidance on the hyperparameters tuning (see Appendix C) for machine learning models are obtained to maximise the accuracy of the models.

*Table 1. Companies in the violation of CSRC Enforcement of Actions*

| Stock | Violation Date | Violation ID | [Other Fields] | Violation or Not? |
|---|---|---|---|---|
| 0001 | 03 May 2020 | P2501 | … | YES |
| 0002 | 20 October 2021 | P2501 | … | NO |
| 0003 | 19 January 2020 | P2502 | … | YES |
| … | … | … | … | … |

## 2.2. Exploratory Data Analysis (EDA)

EDA is an approach, or a set of techniques used to analyse and summarize datasets in order to better understand their main characteristics, patterns, and relationships between variables. One use of EDA in our context is anomalies detection. For example, the Figure 1 below shows the basic summary statistics of P/E Ratio, which is one of our model variables. As we can see from the box plot diagram, there is one record lying above 800 for the index axis, with standard deviation reaching an unreasonable value of 166.62. This implies data removal is required to ensure the robustness of the machine learning models.

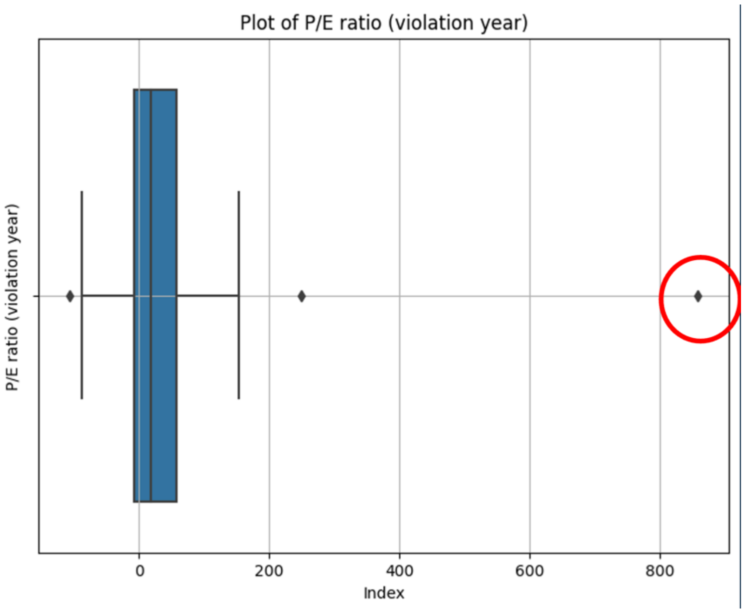


*Figure 1. Box plot for price earnings ratio*

*Table 2. Summary Statistics for price earnings ratio*

| Count | Mean | SD | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|
| 30 | 56.805333 | 166.621461 | -105.59 | -6.3475 | 18.675 | 58.295 | 858.84 |

In addition, we analyse the correlation among the parameter using correlation heatmap. According to the Figure 2 on the right. Letter A stands for the binary output which are "Yes" or "No" for market manipulation. Letter B to I are all the parameters to be included in the machine learning models. We can see Letter A and C have the highest correlation of 0.46. By referring to the definition of C, the Average Volume 60 days before and after the manipulation date is the most correlated column with the Violation. This gives us insight for the stage of parameters tuning.

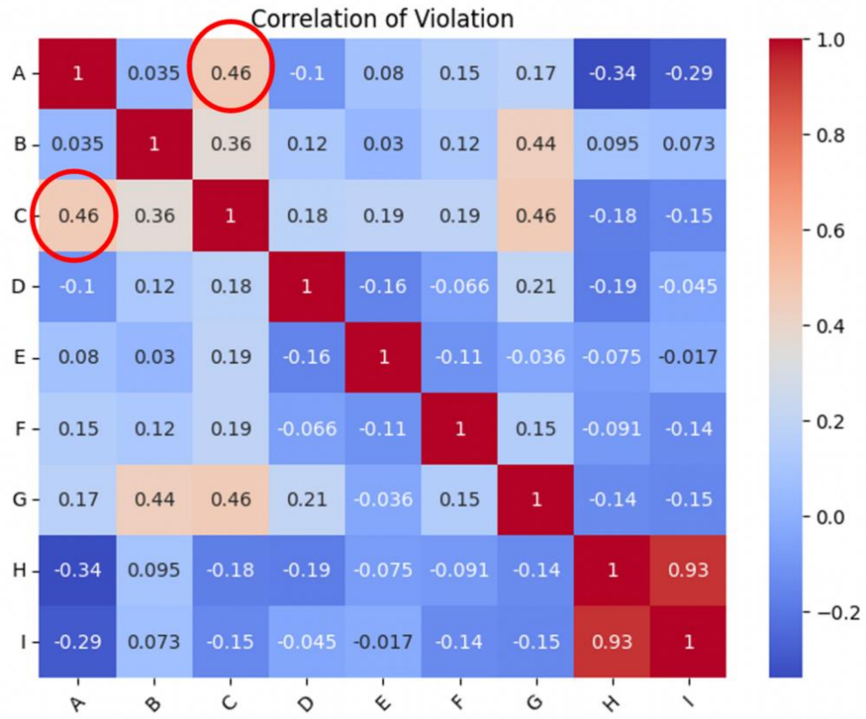


*Figure 2. Correlation plot for the features to the violation*

### 2.3 Data Removal & Adjustment

After exploring the data, we remove and adjust the dataset based on the rationale shown in Table 3 below, before being integrated into the models. Additionally, we standardized the dataset, including prices, to prevent a wide variation of numerical attributes.

*Table 3. Data Cleaning Process*

| Situation | Description |
|---|---|
| Delisted firm. | Remove related violation records. |
| 60 days back-and-forth without trades. | Remove related violation records.<br>(Stocks that are highly illiquid, extremely small market cap or shareholding highly concentrated which the board of management won't trade in open market frequently) |
| Unspecific case categories | Remove related violation records.<br>(There is a generic category call "Others" that might undermine other characters)<br>(Furthermore, some categories are not mutually exclusive) |
| Incidents logged on non-trading days | Take the closest trading day. |
| No data in a concerned period. | Take the average of the start and the end date. Discard violation records with no data for more than 5 trading days consecutively. |

| | |
|---|---|
| No inventory turnover. | Take the parameter as 0 zero, factor weighting turns 0.<br>(Software-as-a-service and agency business sectors do not have inventory turnover) |

### 2.4 Data Processing

As shown in Figure 3 below, this is the visualised dataset with orange squares being samples **with market manipulation** and green squares being samples **without market manipulation**. The final dataset used in the machine learning models comprises 1,500 manipulating companies and 800 non-manipulating companies with complete data. Obviously, the dataset is imbalanced (see left figure). For experimental purpose, we choose 10 cases from each year (5 “Yes” & 5 “No”), which sums to 30 cases in total (see right figure). We can reasonably assume Central Limit Theorem (CLT) holds where the sum or average of a large number of independent and identically distributed (i.i.d.) random variables will be approximately normally distributed, regardless of the distribution of the individual variables. **In the final report, we will for sure scale-up the sample size in a great extent.**

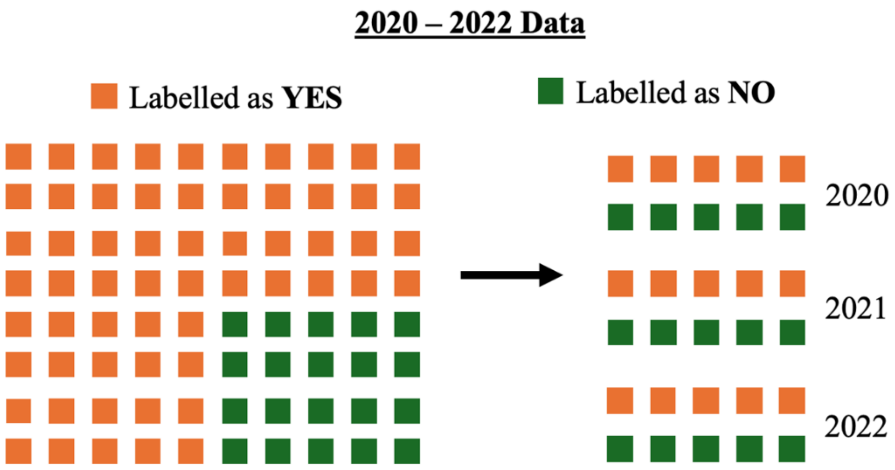


*Figure 3. Current labelled dataset*

### 2.5 Collection of Model Parameter for Each Record

Upon collection of the list of companies involved in the market manipulation, an in-depth analysis of the respective violations was conducted to ascertain the specific manipulation date for each infraction. Subsequently, pertinent trading data, encompassing the average price 60 days before and after the manipulation date, average volume 60 days before and after the manipulation date, inventory turnover ratio, price-earnings ratio, beta, realized volatility, current ratio, and quick ratio on the manipulation date, was gathered as outlined in Table 4. The data's structure and format were standardized across all selected listed companies to ensure consistency in the analysis.

*Table 4.* Features for the machine learning model

| **Features** | **Timeframe** | **Definition** |
|---|---|---|
| Average Price | ±60 days of the manipulation date | $Price\ of\ stock$ |
| Average Volume | ±60 days of the manipulation date | $Number\ of\ shares\ traded$ |
| Inventory turnover ratio | Manipulation year | $\frac{Cost\ of\ Goods\ Sold}{Average\ Price\ of\ Inventory}$ |
| Price earnings ratio | Manipulation date | $\frac{Current\ Market\ Price}{Earning\ per\ Share}$ |
| Beta | Manipulation date | $\frac{Covariance(r_i, r_m)}{Variance(r_m)}$ |
| Realized Volatility | Manipulation date | $Standard\ deviation\ of\ return$ |
| Current Ratio | Manipulation year | $\frac{Current\ Asset}{Current\ Liability}$ |
| Quick Ratio | Manipulation year | $\frac{Cash\ \&\ Cash\ Equivalents}{Current\ Liability}$ |

The features were considered as possible factors to detect market manipulation. First, there will be significant fluctuations in price and trade volume when manipulations occur, so average price and volume are selected. Second, inventory turnover ratio, current ratio, and quick ratio indicate the liquidity of a company, which further shows the ability to sell the company’s stock in the market. Manipulators may attempt to distort a company’s financial statements to mislead other retail investors. By looking at the ratio and comparing it with industry peers, significant deviations from the norm can indicate potential manipulations, warranting further investigations. Third, beta and realized volatility are related to a company’s risk. Higher risk can be associated with higher expected return. High values can make it easier for manipulators to create false and misleading appearances with respect to the price of security. Selection of these features are then justified with the above reasons.

## 2.6 Model Building

The selected machine learning models for detecting market manipulation are as follows: Support Vector Machines (SVM), Decision Tress (DT), Naïve Bayes (NB), and Logistic Regression (LR). These models were chosen for their abilities to handle high-dimensional data and mitigate overfitting (see Appendix D) by minimizing the loss function during training.

### 2.6.1 Support Vector Machines

SVM excels in classification and pattern recognition, particularly for binary categorization of fraudulent and non-fraudulent cases. SVM is effective in high-dimensional spaces but may not be suitable for large datasets with substantial noise.

### 2.6.2 Decision Trees

DT handle both numerical and categorical data by creating a hierarchical structure of decisions based on features. They can capture non-linear relationships in data and perform feature selection. However, pruning is necessary to prevent overfitting.

### 2.6.3 Naïve Bayes

NB is based on Bayes' theorem and assumes the features are conditionally independent given the class label. Despite its simplicity, NB performs well in terms of its training speed and prediction speeds, even with large datasets.

### 2.6.4 Logistic Regression

LR is commonly used for binary classification problems. It interprets feature relationships using a linear equation, capturing linear patterns. While computationally efficient for large datasets, LR may not be suitable for highly complex and non-linear fraud patterns.

These four machine learning models are coded in Python, leveraging its extensive libraries and frameworks for data analysis. For instance, we used NumPy and Pandas for data cleaning, exploration, and preprocessing, and Matplotlib and Seaborn for data visualization to enhance presentation and understanding. Python contributes an impressive degree of potency and adaptability to machine learning ecosystems.

## 2.7 Model Evaluation

To compare the model's performance, we use accuracy, a common metric that assesses overall model effectiveness as shown in (1).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

Sensitivity measures the true positive rate, or the proportion of positive results correctly classified, as indicated in (2).

$$Sensitivity = \frac{TP}{TP+FN} \tag{2}$$

Specificity quantifies the true negative rate, or the proportion of negative results correctly classified, as expressed in (3).

$$Specificity = \frac{TN}{TN+FP} \tag{3}$$

A more comprehensive evaluation includes the AUC-ROC curve, which visualizes the model's performance. The ROC plot showcases the trade-off between true positives and false positives across various threshold values, while AUC, ranging from 0 to 1, measures the model's ability to distinguish between positive and negative classes. Higher AUC values indicate superior discrimination of positive and negative classes.

## SECTION 3 – RESULTS AND FINDINGS

During the initial phase of this project, the primary emphasis was on executing the first four steps outlined in methodology – from data collection to model building. To identify the model that demonstrated the most consistent performance, four machine learning models were selected for training and subsequent comparison. In Section 3.1, it involves the discussion of the results obtained from all four models. Additionally, Section 3.2 delves into the upcoming semester's focus on fine-tuning the hyperparameters of the models to further optimize their effectiveness.

### 3.1 Results and Discussions

During the training process, 30 stock records with 25% testing size were fit into four models respectively. Results are shown in Table 5. Respective ROC curves are shown in Appendix E.

*Table 5.Results of All Four Machine Learning Models in terms of Accuracy and AUC*

| | **Accuracy** | **AUC** |
|---|---|---|
| Support Vector Machine | 0.75 | 0.93 |
| Decision Tree | 0.5 | 0.6 |
| Naïve Bayes | 0.625 | 0.8667 |
| Logistic Regression | 0.75 | 0.73 |

SVM overall has the best performance, with the highest accuracy and AUC. It is considered effective in the current number of features with a small dataset. However, its performance is doubted after the expansion of dataset and features. Since SVM requires a long period of training time, it may be less practical for large datasets.

Naïve Bayes has a relatively good performance, with the second highest accuracy and AUC. From the current results, it is a simple and fast model which is effective at predicting classes, even in a small dataset. However, the impractical underlying assumption of independence among features should not be neglected. Its performance is still possible to be negatively affected after the expansion.

Logistic Regression has a similar performance to Naïve Bayes. It has the highest accuracy and second highest AUC among the four models. The results show that it is effective in detecting market manipulation. Nevertheless, this model may perform poor if the true relationship between independent variables and dependent variables are found not to be linear, after expanding the dimension of features.

Decision tree has the poorest performance, which is like a random guessing algorithm with 0.5 accuracy. Since the decision tree is sensitive to data size, the poor performance may be caused by the insufficient amount of data. With a small amount of data, it may find struggled to analyze patterns and generalize well. However, it is sensitive to variations in training data. When the training data is further enlarged, it may draw to a totally different conclusion.

## 3.2 Hyperparameters Tuning

After expanding the number of stock records and dimension of features, hyperparameters of each model will be tuned to get an optimal solution (shown in Table 6).

*Table 6. Hyperparameter Tuning*

| | **Hyperparameters** | **Purpose** |
|---|---|---|
| SVM | ● Kernel type<br>● Regularization parameter | ● Avoid overfitting |
| DT | ● Maximum depth<br>● Minimum number of samples in leaf node<br>● Criterion for splitting | ● Capture complex relationship<br>● Avoid overfitting<br>● Better feature selection |

| | | |
|---|---|---|
| | ● Implement Adaboost and Random Forest | |
| NB | ● Smoothing parameter<br>● Prior probability of each class | ● Avoid overfitting |
| LR | ● Regularization parameter<br>● Maximum number of iterations | ● Avoid overfitting<br>● Ensure convergence on the optimal solution |

## SECTION 4 – FUTURE POSSBILITIES / PENDING IMPLEMENTATIONS

Firstly, the existing dataset may not adequately showcase the learning capabilities of each learning model. Accordingly, **there are plans to augment the dataset within the same timeframe to demonstrate the performance of each learning model more comprehensively**. The forthcoming interim report will meticulously outline and justify the performance comparisons for each learning model. However, there is a recognition of the potential challenge posed by imbalanced datasets anticipating the expansion of the dataset. Consequently, the SMOTE technique is being considered as one of the potential solutions to address this issue. Results obtained both with and without the application of the SMOTE technique will be systematically tabulated to discern and analyze the differences in results between the two scenarios.

Second, **subsequent project updates will include sample results of key evaluation metrics such as confusion matrix, accuracy, sensitivity, specificity, and AUC for each machine learning model**. Comparative analysis and ROC curves will provide enhanced visualization, facilitating the selection of the optimal model based on comprehensive performance assessments across all models. Additionally, the training and testing error for each iteration will be measured to avoid the possibility of underfitting or overfitting.

Third, market violations such as price manipulations may involve investors disseminating false information to artificially inflate stock prices. Future implementations will **explore the inclusion of social media analysis as input features for machine learning models**. The Chinese stock forum, Guba is considered given the substantial number of retail investors who share their insights and opinions on the Chinese stock market. This presents an opportunity to provide additional information to the models specifically related to manipulated stocks. Acknowledging the potential uncertainties introduced by social media data, the study can further enhance to contemplate the **implementation of sentiment analysis to enhance the robustness of the machine learning model**. Specifically, information extracted from Guba, including the volume of messages related to a particular stock, relevant keywords, and the number of posts read counts, will be subjected to sentiment analysis. This analytical approach aims to identify abnormal sentiment patterns, such as sudden surges in positive or negative sentiment, indicative of a higher likelihood of market manipulation. Additionally, engagement metrics such as read counts, likes, shares, and other interactions will be collected, enabling the model to raise flags in the presence of high engagement with suspicious sentiment during manipulation dates. (see Appendix F)

Fourth, the **machine learning models can be refined to tailoring models to specific types of market manipulations identified in the CSRC's Enforcement Action database**. Rather than a broad flagging of market manipulation occurrences, this targeted strategy involves implementing machine learning models designed to detect violations on a case-by-case basis (see Appendix G). For instance, the SVM may excel in detecting violations associated with ID P2501, while the LR model may be more effective in identifying violations linked to ID P2502. This tailored approach enhances the precision of detection for specific violation cases, offering insights that can potentially reveal unseen patterns and contribute to more nuanced research in this domain.

Fifth, in the subsequent phase of our research, the objective of the project is **to transition from historical data-based detection to real-time monitoring of market manipulation in stocks**. This involves training machine learning models in real-time, utilizing features collected from various sources. By adopting a proactive approach, the models can identify suspicious activities as they occur. This real-time monitoring does not only facilitate swift detection but also enables investigation authorities to promptly address and prevent potential public losses associated with market manipulation.

Spanning from September 2023 to April 2024, the project progresses through distinct phases. As for the current progress, we have just completed our first project presentation and are currently working on the implementation and a detailed interim report. Subsequent months prioritize parameter fine-tuning and feature extraction. April 2024 concludes the project with the final presentation and submission of the comprehensive final report. A detailed schedule breakdown is provided for reference (Table 7).

*Table 7: Proposed Timeline*

| **Timeline** | **Description** |
|---|---|
| Mid Oct 2023 | Communicate with Dr. Liu after submission of project plan |
| Early Nov 2023 | Complete the data collection, cleaning and pre-processing<br>Touch-up on the website for presentations of results |
| End Nov 2023 | Setting up the machine learning models |
| Dec 2023 | Calculations on the hyperparameters for machine learning models |
| 11 Jan 2024 | First presentation |

| | |
|---|---|
| **21 Jan 2024** | **Deliverables of Phase 2 (including Preliminary implementation and Detailed Interim Report)** |
| Jan – Mar 2024 | Final parameters tuning for the most optimal model, advanced feature extraction |
| 15 Apr 2024 | Final presentation |
| 23 Apr 2024 | Deliverables of Phase 3 (Submission of Final Report) |

## SECTION 5 – CONCLUSION

This paper emphasizes the need to safeguard the integrity of investors in financial markets, particularly considering the escalating instances of market manipulation. We aim to introduce an optimal machine learning model for detecting market manipulation in stocks listed on the Shanghai Stock Exchange and Shenzhen Stock Exchange. The evaluation of model performance relies on a comparison of confusion matrices and ROC analysis results from SVM, Naïve Bayes, DT and LR.

Our empirical findings are expected to indicate that the SVM model stands out as the most accurate and consistently efficient tool for detecting market manipulation. The evaluation metrics achieved by SVM will be studied to compare the value surpasses the existing solutions. This can help to enhance investor protection and market integrity. However, there are limitations to be acknowledged. One limitation is the focus on historical data that restricts our ability to detect manipulation in real-time. Future research should explore modifications to enable real-time prediction and proactive identification of market manipulation. This would help mitigate financial losses and safeguard the broader economy. Furthermore, we recommend incorporating sentiment analysis to study external factors that influence market behavior. By eliminating outliers and noise from the model, its performance can be further enhanced, enabling the detection of manipulation under a broader range of conditions. The false positive rate can be greatly reduced in the volatile securities market.

In summary, this paper addresses the critical issue of market manipulation detection and underscores the potential of SVM as an efficient solution. By proactively combating market manipulation, we can better protect the interests of investors and the overall health of the financial markets.

# REFERENCES

Charltons. (2022). Market misconduct under the Securities and Futures Ordinance. 13. Retrieved from https://www.charltonslaw.com/hong-kong-law/market-misconduct-under-the-securities-and-futures-ordinance/

Franck, T., & Son, H. (2021). JPMorgan hit with $200 million in fines for letting employees use WhatsApp to evade regulators’ reach. *CNBC*. Retrieved from https://www.cnbc.com/2021/12/17/jpmorgan-agrees-to-125-million-fine-for-letting-employees-use-whatsapp-to-evade-regulators.html

Gantz, R. (9 6, 2022). *When Markets Are Bearish, Beware Of Stock Manipulation*. Retrieved from Forbes: https://www.forbes.com/sites/forbesbusinesscouncil/2022/06/09/when-markets-are-bearish-beware-of-stock-manipulation/?sh=1fb600e33b00

Goel, A. (2023). The relationship between economy & stock market. Retrieved from https://economictimes.indiatimes.com/markets/stocks/news/the-relationship-between-economy-stock-market/articleshow/96878611.cms

JPX. (2023). *Market Manipulation*. Retrieved from Japan Exchange Group: https://www.jpx.co.jp/english/regulation/preventing/manipulation/index.html

Kottasova, I. (2016). Financial fraud happens every 15 seconds. *CNN Money*. Retrieved from https://money.cnn.com/2016/09/20/news/financial-fraud-every-15-seconds/index.html

Li, A., Wu, J., & Liu, Z. (2017). *Market Manipulation Detection Based on Classification Methods.* doi:https://doi.org/10.1016/j.procs.2017.11.438

Liu et al. (2021). *Detecting stock market manipulation via machine learning: Evidence from China Securities Regulatory Commission punishment cases.* doi:https://doi.org/10.1016/j.irfa.2021.101887

Livni, E. (2023). Just How Common Is Corporate Fraud? *The DealBook Newsletter*. Retrieved from https://www.nytimes.com/2023/01/14/business/dealbook/how-common-is-corporate-fraud.html

Tramplin, T. (2023). *Market Manipulation*. Retrieved from Finance Strategists: https://www.financestrategists.com/financial-advisor/business-ethics/market-manipulation/

Wildau, G. (2017). China securities regulator breaks 2016 record for stock fines. Retrieved from https://www.ft.com/content/d61ab55c-07d5-11e7-97d1-5e720a26771b

Yi, Z., Cao, X., Pu, X., Wu, Y., Chen, Z., Francis, A., . . . Khan, A. T. (2023). *Fraud detection in capital markets: A novel machine learning approac.* Expert Systems With Applications. doi:https://doi.org/10.1016/j.eswa.2023.120760

# APPENDICES

Appendix A

Types and values of a particular company listed in CSRC's Enforcement of Action

| Column | Description | Attribute |
|---|---|---|
| Date | Trading date | Ordinal |
| Open | Daily open price | Numeric |
| High | Daily highest price | Numeric |
| Low | Daily lowest price | Numeric |
| Close | Daily close price | Numeric |
| Adj Close | Daily adjusted closing price | Numeric |
| Volume | Daily Trading Volume | Numeric |

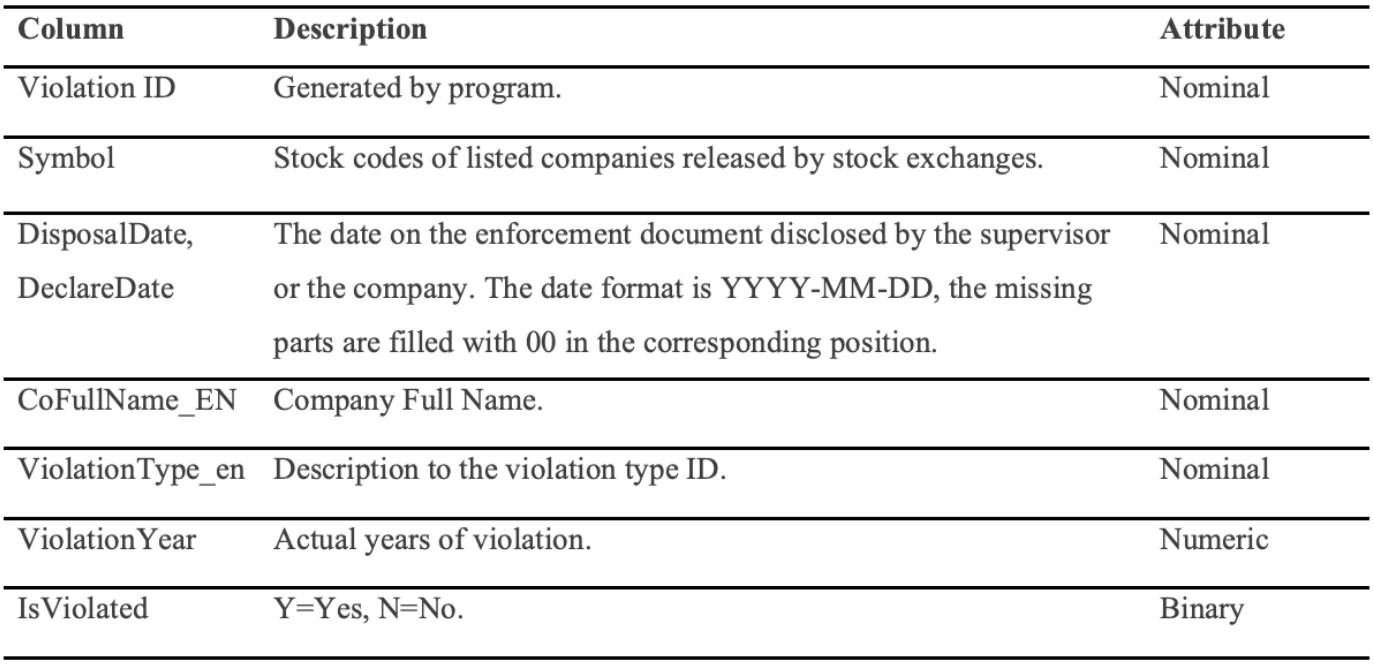
Appendix B

Descriptions for common types of market manipulations (Tramplin, 2023)

| Type | Description |
|---|---|
| Pump and Dump | Artificially boost the price of a security by disseminating false or deceptive information |
| Spoofing | Make fake orders in the market without execution to create a false image |
| Wash Trading | Buying and selling same securities at one time and create an illusion of increased trading volume |
| Insider Trading | Individuals access to non-disclosure trading information, leaving unfair advantage to other investors |
| Cornering the Market | Dominant in a security, commodity, or any financial instrument to manipulate and control the price and supply |
| Front-Running | Exploiting advanced knowledge of impending orders or trades and earning from price fluctuations |

Appendix C

Types and values of each attribute from CSRC’s Enforcement of Action



| Column | Description | Attribute |
|---|---|---|
| Violation ID | Generated by program. | Nominal |
| Symbol | Stock codes of listed companies released by stock exchanges. | Nominal |
| DisposalDate,<br>DeclareDate | The date on the enforcement document disclosed by the supervisor or the company. The date format is YYYY-MM-DD, the missing parts are filled with 00 in the corresponding position. | Nominal |
| CoFullName_EN | Company Full Name. | Nominal |
| ViolationType_en | Description to the violation type ID. | Nominal |
| ViolationYear | Actual years of violation. | Numeric |
| IsViolated | Y=Yes, N=No. | Binary |

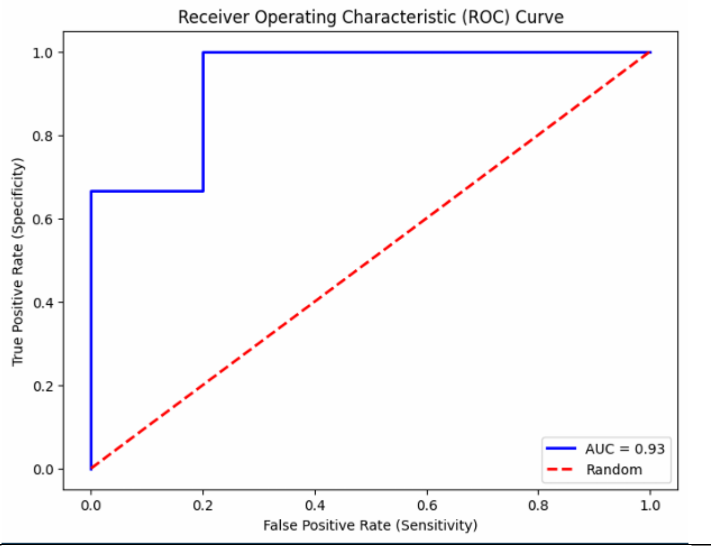
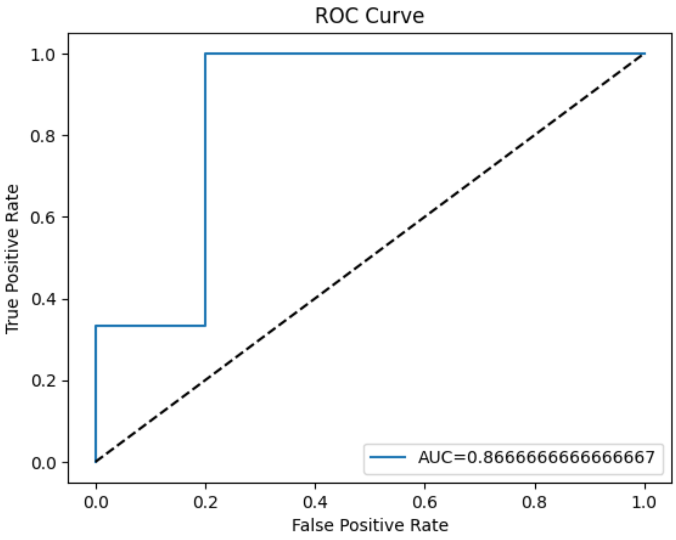
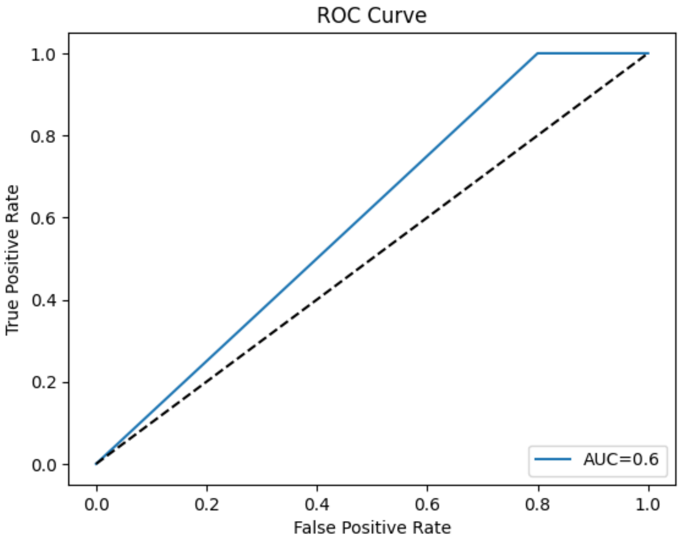
Appendix D

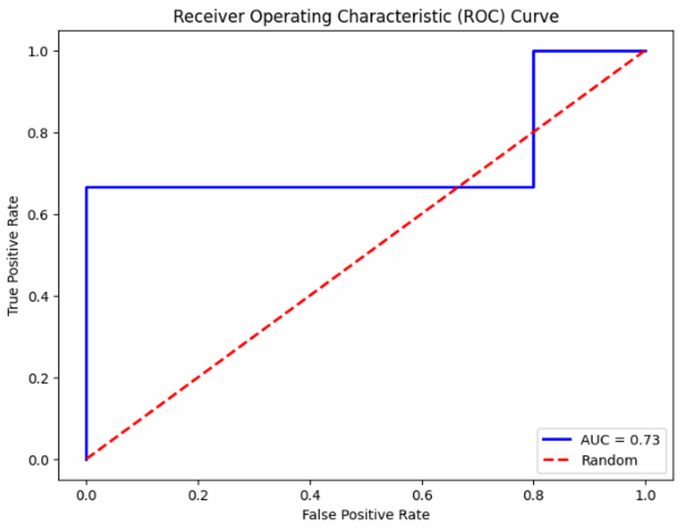
Descriptions for underfitting and overfitting in machine learning

| **Underfitting** | **Characteristics** | **Overfitting** |
|---|---|---|
| Model is not complex | **Model** | Model is too complex |
| Not Accurate | **Training Dataset** | Accurate |
| Not Accurate | **Testing Dataset** | Not Accurate |
| Increase number of features<br>Increase training duration<br>Increase model complexity<br>Remove noise from data | **Reduction Techniques** | Reduce number of features<br>Introduce early stopping<br>Reduce model complexity<br>Increase training data |

Appendix E

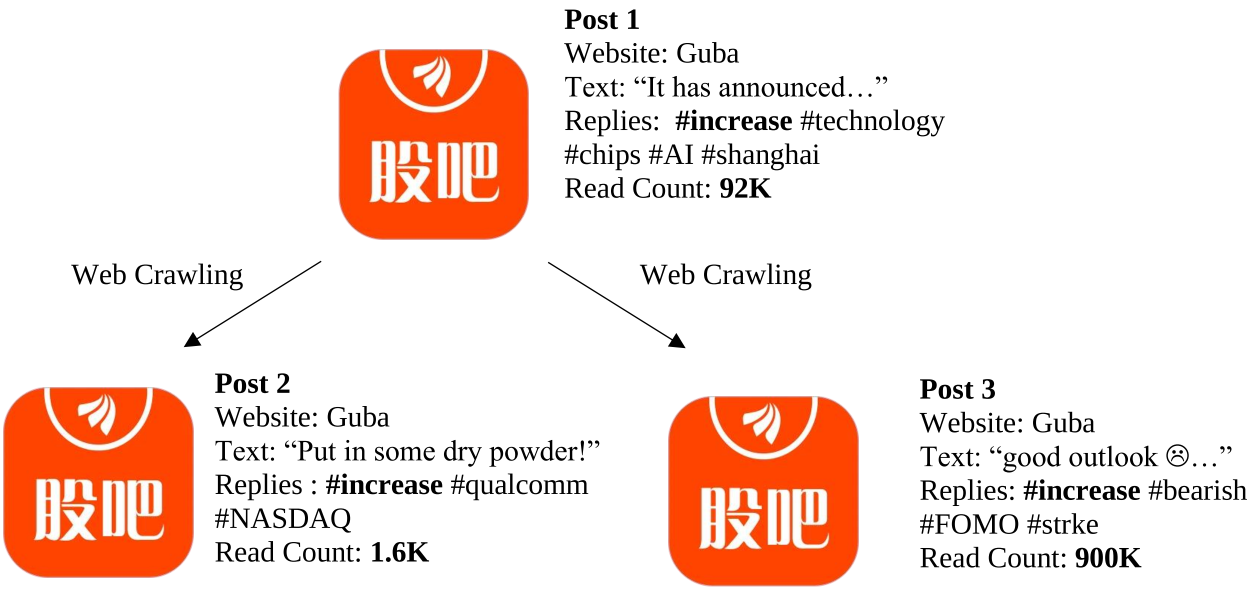
ROC curves of various machine learning models under testing

| **Model** | **ROC curve** |
|---|---|
| Support Vector Machine |  |
| Naïve Bayes |  |
| Decision Tree |  |

| | |
|---|---|
| Logistic Regression |  |

Appendix F

Sentiment analysis for Guba



**Post 1**
Website: Guba
Text: “It has announced…”
Replies:  **#increase** #technology #chips #AI #shanghai
Read Count: **92K**

Web Crawling

Web Crawling

**Post 2**
Website: Guba
Text: “Put in some dry powder!”
Replies : **#increase** #qualcomm #NASDAQ
Read Count: **1.6K**

**Post 3**
Website: Guba
Text: “good outlook ☹…”
Replies: **#increase** #bearish #FOMO #strke
Read Count: **900K**

Appendix G

The types of violation cases with the violation ID

| Violation ID | Violation Cases |
|---|---|
| P2501 | Fictitious Profit |
| P2502 | Fictitious Assets |
| P2503 | False Recordation (Misleading Statements) |
| P2504 | Delayed Disclosure |
| P2505 | Material Omission |
| P2506 | Other False Information Disclosure |
| P2507 | Fraudulent Listing |
| P2508 | False Capital Contribution |
| P2509 | Unauthorized Changes in Capital Usage |
| P2510 | Occupancy of Company’s Assets |
| P2511 | Insider Trading |
| P2512 | Illegal Stock Trading |
| P2513 | Stock Price Manipulation |
| P2514 | Illegal Guarantee |
| P2515 | Mishandling of General Accounting |
| P2516 | Tax Dodging |
| P2517 | Evasion of Tax Arrears Recovery |
| P2518 | Defraud of Export Tax Rebates |
| P2519 | Tax Resistance |
| P2520 | False VAT Invoice or False Other Invoices for Defraud of Export Tax Rebates and Tax Credit |
| P2521 | False Ordinary Invoice |
| P2522 | Printing, Forging and Altering Invoice, Illegal Manufacturing of Anti-counterfeiting Products for Invoice, Forging Supervision Seal for Invoice |
| P2523 | Confirmed by Tax Authority to be Escaped (Lost of Contact) with Behaviours of Tax Dodging, Evasion of Tax Arrears Recovery, Defraud of Export Tax Rebates, Tax Resistance, False Invoice, etc. |
| P2524 | Unpaid or Underpaid Taxes |
| P2599 | Others |