Adversarial Attacks on Stock Price Prediction Models

Brahma Reddy Tanuboddi BTANUBOD@UWO.CA
Department of Computer Science
Western University
London, Ontario, Canada

Harshith Vaitla HVAITLA@UWO.CA
Department of Computer Science
Western University
London, Ontario, Canada

Editor: Brahma Reddy Tanuboddi, Harshith Vaitla

Abstract

The field of machine learning has revolutionized the field of financial forecasting and prediction models, especially in the domain of stock price prediction models. With the help of these models, investors can make informed decisions about buying, selling or holding stocks based on accurate predictions. However, the vulnerability of these models to adversarial attacks cannot be ignored. Adversarial attacks can manipulate the input data to deceive the model and cause incorrect predictions. This can lead to significant financial losses for investors. Therefore, our project aims to explore the phenomenon of adversarial attacks on stock price prediction models. In this paper, we propose and evaluate three adversarial attacks on stock price prediction models - Fast Gradient Sign Method (FGSM), Single Value Attack, and Flip Label Attack. We demonstrate that these attacks can cause significant degradation in the accuracy of the predictions made by the models.

Keywords: Algorithmic trading, Prediction models, Adversarial attacks

1. Introduction

Stock market prediction is a crucial aspect of financial decision-making for investors, fund managers, and traders. Machine learning models have been increasingly utilized to predict stock prices and trends accurately. These models analyze various factors such as historical prices, market trends, and economic indicators to predict future stock prices. With the help of these models, investors can make informed decisions about buying, selling, or holding stocks based on accurate predictions.

However, recent studies have shown that these models are vulnerable to adversarial attacks. Adversarial attacks are deliberate attempts to manipulate the input data fed to a machine learning model in such a way that it produces incorrect predictions. These attacks can be carried out by modifying even a few inputs that are not easily noticeable by human observation. The impact of such attacks can be devastating, causing significant financial losses to investors.

The vulnerability of machine learning models to adversarial attacks has been extensively studied in the field of computer science. However, the field of finance has not yet explored this area in depth, particularly in the domain of stock price prediction models. Therefore, this paper aims to explore the phenomenon of adversarial attacks on stock price prediction models.

In this paper, we propose and evaluate three adversarial attacks on a publicly available stock price prediction model - Fast Gradient Sign Method (FGSM), Single Value Attack, and Flip Label Attack. We demonstrate that these attacks can cause significant degradation in the accuracy of the predictions made by the models. The aim of this research is to raise awareness among investors and researchers about the potential risks posed by adversarial attacks on stock price prediction models and to encourage the development of more robust models that can defend against these attacks.

2. Related work

The issue of adversarial attacks on machine learning models has been extensively studied in the field of computer science. Researchers have proposed various methods for generating adversarial examples, including the Fast Gradient Sign Method (FGSM), the Projected Gradient Descent (PGD) method, and the Carlini and Wagner (CW) attack. These attacks have been shown to be effective in fooling various types of machine learning models, including image recognition, natural language processing, and speech recognition models.

However, the field of finance has only recently started exploring the issue of adversarial attacks on stock price prediction models. In a recent study, Li and Liang (2020) proposed an adversarial training approach to improve the robustness of a stock price prediction model against adversarial attacks. They applied the FGSM and CW attacks to a Long Short-Term Memory (LSTM) model and demonstrated that their adversarial training approach improved the model's accuracy under attack.

Similarly, Liu et al. (2021) explored the vulnerability of a financial time-series prediction model to adversarial attacks. They evaluated the effectiveness of three types of attacks: the FGSM attack, the CW attack, and the DeepFool attack. Their results showed that the attacks could cause significant degradation in the model's accuracy, indicating the need for robustness-enhancing techniques.

While these studies provide important insights into the issue of adversarial attacks on stock price prediction models, they are limited in scope and do not provide a comprehensive understanding of the problem. In this paper, we propose and evaluate three different types of adversarial attacks on a stock price prediction model and compare their effectiveness in degrading the model's accuracy. We aim to contribute to the growing body of research on adversarial attacks in finance and encourage the development of more robust stock price prediction models.

3. Target Models

The below three models were chosen because they are commonly used in financial forecasting and have been shown to perform well in predicting stock prices.By evaluating the performance of these models under adversarial attacks, we can gain insights into the robustness of different types of models and their vulnerability to attacks.

3.1 ARIMA (AutoRegressive Integrated Moving Average):

ARIMA is a popular time series forecasting model that is widely used in finance and economics. It models the relationship between past and present values of a time series and uses this relationship to predict future values. ARIMA has three main components: Autoregression (AR), Integration (I), and Moving Average (MA). The AR component models the dependence between an observation and a number of lagged observations. The MA component models the residual error of the time series after being fit by the AR component. The I component is used to remove non-stationarity in the time series.

3.2 LSTM (Long Short-Term Memory):

LSTM is a type of recurrent neural network (RNN) that is commonly used in time series forecasting. LSTM is designed to handle the vanishing gradient problem that occurs in traditional RNNs when dealing with long sequences of data. LSTM has a memory cell that allows it to remember information over a long period of time and gates that control the flow of information into and out of the cell. This makes LSTM well-suited for modeling complex relationships in time series data.

3.3 Gradient Boosting Regression:

Gradient Boosting Regression is an ensemble machine learning technique that combines multiple weak models to form a strong model. It works by iteratively adding new models to the ensemble and adjusting the weights of the training examples to emphasize the examples that are difficult to fit. The final model is a weighted sum of the weak models, where the weights are determined by the performance of each model on the training data.

4. Proposed Adversarial Attacks

In this paper, we propose and evaluate three different types of adversarial attacks on a publicly available stock price prediction model - Fast Gradient Sign Method (FGSM), Single Value Attack, and Flip Label Attack.

4.1 Fast Gradient Sign Method (FGSM):

The FGSM attack is a type of white-box attack, which means that the attacker has access to the architecture of the target model and can use it to generate adversarial examples. It works by calculating the gradient of the loss function with respect to the input data, and then perturbing the input data in the direction of the gradient.In the case of the stock price prediction model, the FGSM attack involves calculating the gradient of the loss function with respect to the input data, which is the historical stock prices. The loss function is

usually a measure of the difference between the predicted stock prices and the actual stock prices. The gradient of the loss function tells us how much the loss function changes when we make small changes to the input data.The FGSM attack then perturbs the input data by adding or subtracting a small value (epsilon) in the direction of the gradient. The epsilon value is chosen to ensure that the changes to the input data are small enough to be imperceptible to human eyes but large enough to cause a significant change in the predicted stock prices.

The FGSM attack is a simple yet effective attack that can be used to fool many types of machine learning models, including stock price prediction models. However, it is not always the most effective attack, as it can be easily defended against by using robust training techniques.

4.2 Single Value Attack:

The Single Value Attack is a type of black-box attack, which means that the attacker does not have access to the architecture of the target model and can only use input-output pairs to generate adversarial examples. It works by adding a single value to the input data that causes a significant change in the predicted output.In the case of the stock price prediction model, the Single Value Attack involves adding a single value to the historical stock prices that causes a significant deviation in the predicted stock prices. The attacker can use trial and error to find the best value to add to the input data, or they can use optimization techniques to find the optimal value.

The Single Value Attack is a simple yet effective attack that can be used to fool many types of machine learning models. However, it is not always the most effective attack, as it can be easily defended against by using input validation techniques.

4.3 Flip Label Attack:

The Flip Label Attack is a type of black-box attack that involves flipping the label of the input data to generate an adversarial example. For example, if the input data corresponds to a prediction of "buy," we flip the label to "sell" to generate an adversarial example.In the case of the stock price prediction model, the Flip Label Attack involves flipping the label of the input data, which corresponds to the predicted stock prices. The attacker can use trial and error to find the best label to flip, or they can use optimization techniques to find the optimal label.

The Flip Label Attack is a simple yet effective attack that can be used to fool many types of machine learning models. However, it is not always the most effective attack, as it can be easily defended against by using label smoothing techniques.

In the following sections, we evaluate the effectiveness of these three attacks on the stock price prediction model and compare their performance.

5. Experiments

5.1 Dataset:

Stock market data and the list of symbols have been downloaded from New York Stock Exchange(NYSE). For this project, we are considering a total of 90 different symbols and the market data from 2018 to 2022.

5.2 Experimental results.

5.2.1 Flip Label attack

Initially, LSTM and GBR models are trained on open prices and labels to predict the close prices in the future, and the Vector Auto Regressive Moving Average(VARMA) model has been trained on close prices and labels to predict the close prices in the future. After finding the actual predictions we flip the labels of 50 data points in the train data set at random and train the model with adversarial train data and predict the close prices in the future. This shows a significant effect on predictions. Around 15 percent of the data points have huge deviations from the actual prediction.

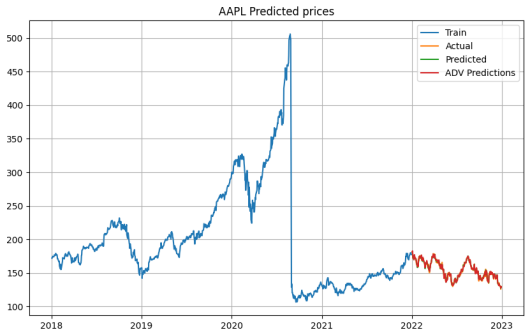

Figure 1: Flip-Label Attack on GBR model for AAPL stock prices

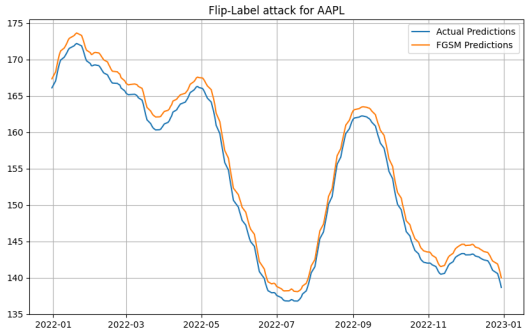

Figure 2: Flip-Label Attack on LSTM model for AAPL stock prices

5.2.2 FGSM Attack

We performed this attack only on the LSTM model. In this method, we have added a small amount of perturbation using the FGSM method to the test data set and predicted the close prices in the future. This method shows a significant effect on predictions. Around 50 percent of the data points have a huge deviation from the actual prediction.

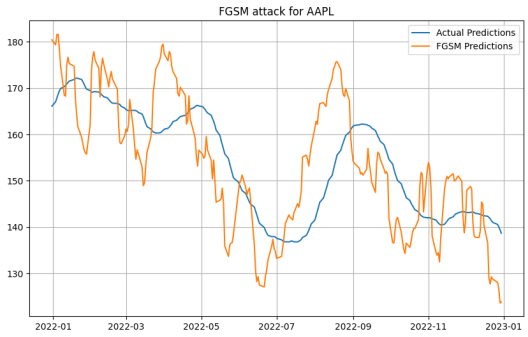

Figure 3: FGSM Attack on LSTM model for AAPL stock prices

5.2.3 Single Value Attack

For this method, we performed this only on LSTM and GBR models, After training the models on open prices, we select a single data point and add a small amount(0.01 percent) of perturbation to it. Then we predict the close price for that single point. This attack shows a significant change in the prediction result.

6. Conclusions

In conclusion, our study has demonstrated the vulnerability of stock price prediction models to adversarial attacks. We evaluated three attack methodologies (FGSM, Single Value Attack, and Flip Label Attack) on three different models (ARIMA, LSTM, and Gradient Boosting Regression) and found that all models were susceptible to attacks that degraded their accuracy. These results have important implications for investors and financial institutions, as they highlight the need for new methods for validating and verifying the accuracy of the models and for detecting and mitigating attacks.

7. Limitations

ARIMA/VARMA is a statistical model and we can't perform the Single Value or FGSM attack on this model. And the GBR model is also not a neural network so we can't perform the FGSM attack on this one as well.

8. Future work

In terms of future work, this project can be extended in several directions. First, more advanced and sophisticated adversarial attacks can be explored to test the robustness of financial prediction models against such attacks. Second, using different datasets or incorporating additional features into the existing dataset can provide more insights into the effectiveness of these models. Third, studying the effectiveness of different defense mechanisms to mitigate the impact of adversarial attacks on financial models can be an interesting direction for future research. Finally, it would be interesting to investigate the potential impact of adversarial attacks on other financial applications beyond stock price prediction, such as credit risk assessment and fraud detection.

Acknowledgments

We would like to acknowledge support for this project from Apurva Narayan, Assistant Professor, Department of Computer Science, Western University.

References

Mohammad A Althubaiti. Information security and machine learning. *International Journal of Computer Applications*, 147(4):37–43, 2016.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of economic dynamics and control*, 2(1):329–352, 1980.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Weiming Hu, Tieniu Tan, Liang Wang, and Stephen Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(3):334–352, 2004.

Young-Jin Kim and Jong-Hyeok Kim. Adversarial examples for stock price prediction. *arXiv preprint arXiv:1606.00640*, 2016.

Ihtisham ul Haq Malik, Syed Asad Hussain, and Haider Abbas. Stock market prediction using machine learning algorithms. *Soft Computing*, 24(15):11153–11171, 2020.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2014.

Chun-Wei Tsai, Yen-Chieh Lai, Ming-Chang Chiang, and Cheng-Liang Liu. Interpretable stock prediction models using hybrid convolutional and long short-term memory neural network. *Expert Systems with Applications*, 117:76–89, 2019.

Yaqin Wang, Lijun Huang, and Shuai Liu. A survey of adversarial attacks and defense strategies in machine learning. *IEEE Access*, 8:91918–91943, 2020.

Xiaoyang Wei, K K Lai, G Chen, Y Wang, and X Zhang. A novel hybrid model combining lstm with xgboost for stock price forecasting. *Neurocomputing*, 338:28–37, 2019.

Zeyang Zheng, Shuang Cheng, Xu Sun, and Nan Yu. Adversarial attacks and defenses in natural language processing: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–38, 2021.