FEATURE ENGINEERING OF UNSUPERVISED STOCK MANIPULATION DETECTION MODEL

CONTROL SYSTEMS AND INSTRUMENTATION ENGINEERING PROGRAM

Kamonpat Thrupmonchai , Korim Akcaramanat , Advisor : Assoc.Prof.Dr.Poj Tangamchit

Introduction

Stock price manipulation harms all investors and reduces trust in financial markets. As manipulators use more advanced methods, including AI, human oversight at the SET is no longer enough. This research builds on previous LSTM Autoencoder models by introducing a **Transformer Autoencoder with new feature engineering** on price data in the form of a limited order book. The improved model detects abnormal trading more effectively and helps promote fairer, more transparent markets.

Method

This project develops a Transformer Autoencoder using Unsupervised Learning to identify abnormal trading patterns. The model is capable of detecting potential manipulation without labeled data, making it effective against evolving trading strategies.

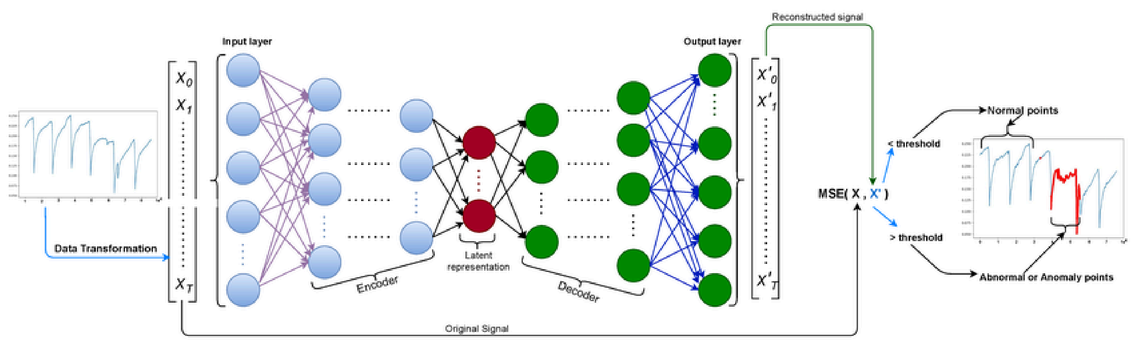


Fig.1 *Anomaly detection pipeline using deep autoencoder-based neural network model*

The Transformer Autoencoder adapts the Transformer architecture for time-series financial data to detect stock price manipulation. Combining the Transformer framework with an Autoencoder, it learns and reconstructs normal trading patterns, using reconstruction errors to flag potential manipulation. Fig. 2 shows the model structure.

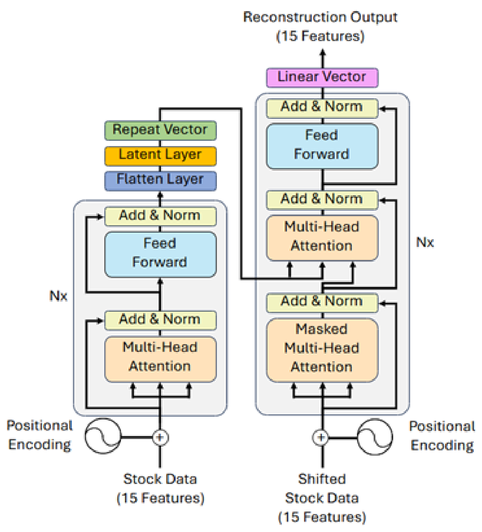


Fig.2 *Transformer Autoencoder Structure*

The model was trained on normal trading data in limited order book format from the Stock Exchange of Thailand. After tuning, it demonstrated improved performance over the previous LSTM Autoencoder model.

References

[1] T. Leangarun, P. Tangamchit, S. Thajchayapong “Stock Price Manipulation Detection Using Deep Unsupervised Learning: The Case of Thailand”, IEEE Access, Vol.9.

Results

The first is the result made by comparing two model detection ratios. Let us first understand the method of calculation. The detection ratio is calculated by
(Inm , Imm / length) * 100
Inm = Normal detect as manipulated , Imm = Manipulate detect as manipulated

Stock	15 Features	34 Features
ASIAN	0	22.865×10^{-2}
FLOYD	0	242.244×10^{-2}
GSC	852.684×10^{-2}	7087.505×10^{-2}
MBAX	183.789×10^{-2}	181.683×10^{-2}
RP	0	857.022×10^{-2}
SCI	0.772×10^{-2}	8.858×10^{-2}

Table.1 *Detection ratio of the model*

From the table we can clearly see those 34 features shadowed 15 features by a whole lot. But to truly understand why we will have to look at the graph below which we will get into shortly. The reason for the number 0 appearing mean that's the model cannot detect manipulation in the dataset provided and the higher the number the better.

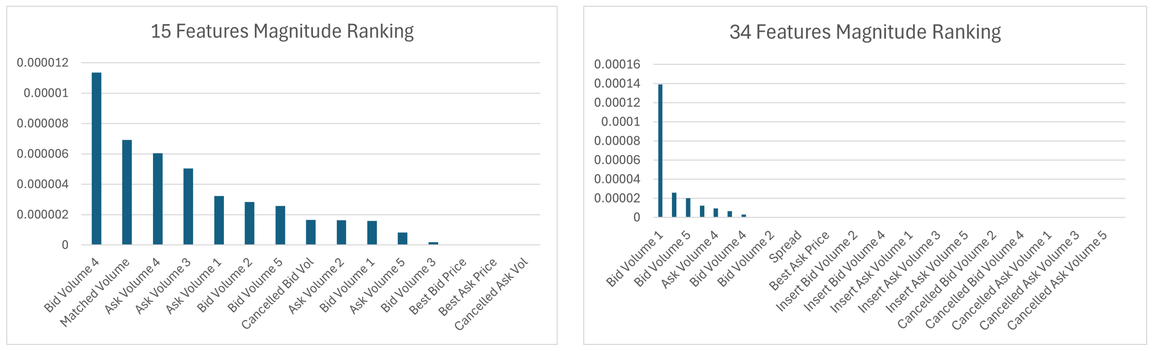


Fig.3 and Fig.4 *Magnitude Ranking*

Now the part of feature importance. In this experiment we have decided to use SHAP features which measure the decision magnitude of each feature and output them as SHAP values. This is what we will call magnitude. We will then rank them based on the contribution to the final answer. The more value of magnitude the better.

Conclusion

From the first table we can clearly see the difference between 15 features and 34 features model. The 34 features model out-performed 15 features, and the score is much better. There's one outstanding model which was GSC dataset. When we take a deeper look at the decision making of both models. The similarity between the main features which contribute to the final decision is found in both models. What this means is that most of the main features are already in the old 15 features model. The reason for the 34 features model is better is because on some occasion features of the 34 features model such as spread and insert volume contribute to the betterment of the model. But on the graph the main contributor is still the old features of the 15 features model.