Insider Trading Detection with Semi-Supervised Learning and Hard-Example Mining

Samuel Burgess
Computer Science
Western Washington University
Bellingham WA, U.S.
saburgess239@gmail.com

Alexander Fortescue
Computer Science
Western Washington University
Bellingham WA, U.S.
fortesa@wwu.edu

***Abstract*—Illegal Insider Trading involves the trading of a companies securities based on non-public information. In the United States, various bodies are involved in the detection of Illegal insider trading including the SEC and FBI. In this report, we detail our work detecting Insider Trading using a supervised classifier that can analyse a window of stock data and output a measure of confidence in whether illegal insider trading was occurring or not.**

***Index Terms*—Deep Learning, LSTM, RNN**

I. Overview

A Financial Security is any financial asset that can be traded. For the sake of this report, a security will refer to a share of a companies stock or some derivative of a companies stock such as a stock option. Insider Trading is technically defined as any trading of a company's securities based on non-public information that could have an effect on that security's price in the future. This project focuses on Illegal Insider Trading in the United States, and as such we will be using data from United States based stock exchanges, namely the New York Stock Exchange (NYSE) and the NASDAQ. In the United States, an 'insider' is defined as someone who is a corporate director, officer, or a stockholder with a stake in a company larger than 10% of the companies total stock. Any trading based on non-public information by these 'insiders' or persons who have been informed by these insiders is illegal under the 1934 and 1984 Securities Exchange acts. Our project seeks to be able to classify some period of a companies stock data history as having been influenced by significant Illegal Insider Trading.

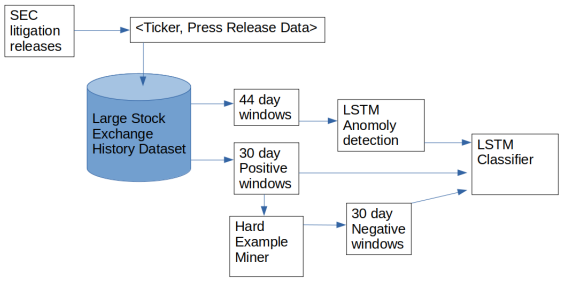
II. Background

Significant Insider Trading generally happens in some time period leading up to the non-public information that is being traded around becoming public. This happens in some form of 'press release' by the company, often an earnings report or an announcement of a merger or acquisition. We use the dates of these press releases as a time to arrange our windows of data around. The volume of a security is the amount of that individual security that is traded on any given day. If 20,000 shares of Alphabet's stock is traded on June 1st, then Alphabet's volume for June 1st is 20,000. We always see a large spike in volume on the day of or few days after the prior mentioned press release. This is due to individuals making new trades to change their position in a company based on this newly available information. Volume is likely the most powerful indicator of Illegal Insider Trading. We are confident that Illegal Insider Trading can be characterized by certain patterns in volume prior to a press release and the associated Volume Spike

III. Prior Work

The prior work most analogous to this project is Rabul Islam et. al's *Mining Illegal Insider Trading of Stocks: A Proactive Approach*. This project involves using an LSTM RNN on time series stock data to analyze cases of Insider Trading. We were able to use a similar data collection method to this paper in order to parse our input data. Islam et. al's paper in turn is largely inspired by Diaz et. al's *Detecting stock market manipulation using supervised learning algorithms*. This paper is more broadly generalized to financial fraud in general, and helped us to conceptualize the anomoly detection portion of our project. The semi-supervised structure of our model(s) is inspired by Carcillo et. al's *Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection*. Which details a number of semi-supervised structures for numerical time series data, and helped us decide on an approach of "best-of-both-world" where an outlier score from an unsupervised model is used to augment the input of a supervised model. We referenced various papers on time series analysis for high level inspiration.

IV. Methodology



We approached our problem with a binary classifier for deciding if a stock window is a case of insider trading. We created this model with an LSTM optimized with adam, using

cross entropy loss. We trained this model on stock data from Kagel's database of all the US-stock market stocks. Kagel's data for each stock includes: the ticker of the stock, the date of trading, the open price, the close price, the adjusted close price, and the volume. This data was then split into windows of 30 days. All of these number values were normalized using z-score normalization to allow more accurate predictions. The windows were assigned a positive feature if they matched the dates from a known case of insider trading from the SEC's website. At first, all other windows were considered to be negative, but with the huge amount of stock data we have access to, this was not the optimal final solution, which we improve with hard example mining.

A. Hard Example Mining

We have access to a few hundred known positive cases, and several hundred thousand assumed negative windows. To make our negative data more useful to the model we decided that we should look through these negative windows and find ones that look similar to the positive cases, so our model will be learning more, from the negative cases. Our plan was, if our model is able to detect small nuances between similar looking windows, instead of just learning what a spike in volume looks like, it will be more more effective at detecting insider trading. We achieved this by sifting through the negative data and creating our negative windows from volume spikes that were of similar size to the positive cases volume spikes. This way the negative cases are not just of normal data, but of chunks of time that had a likely press release or release of information that greatly affected the stock market, and is something someone with insider trading could have taken advantage of before it happened. This causes our model to be able to learn the important nuances we want it to be able to detect.

B. Semi-Supervised Approach

Our less step to making the optimal model came with combining an unsupervised anomaly detection score with our supervised model. We created another LSTM that took the same data, and made windows of 44 days based on every possible volume spike it could find. This time our LSTM was attempting to predict the 14 days prior to the volume spike, based on the 30 days after the spike. We achieved this by using a sliding window that would slide left after each prediction. The LSTM would be fed the first 30 day window, and predict one day before. This prediction was pre-pended to the window, and the final day from the window was removed, creating a new 30 day window to be re-fed to the LSTM to allow the prediction of the day before. This process was done 14 times to create the prediction for the 14 days prior to the spike. This sliding window approach was used because the model is much more accurate when prediction one day before, than predicting farther out. This predicted 14 day window was then compared to the real 14 day window, to give the window an anomaly score. Since insider trading is vary rare, we can assume that of all volume spikes, the most rare cases will be insider trading, and because of this, the highest anomaly scores given are cases we are relatively sure are positive cases, and the less high, the more sure we are the case is a negative case. With this new information, we are able to improve the way our model is adjusting weights when predicting if a case is positive or negative, to result in a more accurate model.

V. Experimental Results

A. Supervised Classifier

All results in this section will be based off a positive-negative window input ratio of 1:2. This is a tunable parameter in our model and currently has a large effect on accuracy. We have found that using a learning rate of 0.001 gives the model 90% accuracy on a test set, classifying 30 day windows as Illegal Insider Trading occurring or not.

B. Unsupervised Anomaly Detection

The Anomaly detection model is able to output a Mean Squared loss of 0.5 between real data and the model's predicted data for the 14 day windows prior to a press release.

VI. Conclusions and Future Work

We believe that our model shows that Time-Series Stock data contains patterns that are indicative or Illegal Insider Trading occurring or not. We have not yet fed the output of our unsupervised model into our supervised model. We will implement this in the coming weeks. We will continue to tune our current model in order to gain greater accuracy using our current methods. In addition we wish to complete the following extensions:

A. Characterize and Visualize Volume Patterns

We hope to be able to empirically define what patterns in stock data is indicative of Illegal Insider Trading. This would allow faster detection of insider trading that does not have to wait until the profits have been made to detect Illegal Insider Trading.

B. Larger Input Size

The major constriction of this project has been the small number of positive windows (known Illegal Insider Trading) of stock data available. A larger labelling operation, access to more SEC records, or natural language processing of existing reports or media could provide a much larger input.

C. Derivative Trading Analysis

Insider Trading is often committed by individuals trading derivatives of a companies stock as opposed to the stock itself. This usually takes the form of options trade, a kind of financial transaction that allows individuals to take massive monetary risk for equally large pay-offs based on whether that individual suspects the stock price will rise or fall. We believe that applying a similar model to option data, if enough is publicly available, would yield great results.

ACKNOWLEDGMENT

We would like to thank Dr. Brian Hutchinson for his guidance and tuition. We would also like to thank Yunjey Choi for their excellent PyTorch tutorials which we have referenced repeatedly.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [**?**].

REFERENCES

[1] Carcillo, Fabrizio Le Borgne, Yann-Aël Caelen, Olivier Kessaci, Yacine Oblé, Frédéric Bontempi, Gianluca. (2019). Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection. Information Sciences. 10.1016/j.ins.2019.05.042.

[2] S. R. Islam, S. Khaled Ghafoor and W. Eberle, "Mining Illegal Insider Trading of Stocks: A Proactive Approach," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 1397-1406.

[3] K. Golmohammadi, O. R. Zaiane and D. Díaz, "Detecting stock market manipulation using supervised learning algorithms," 2014 International Conference on Data Science and Advanced Analytics (DSAA), Shanghai, 2014, pp. 435-441.

[4] Liu, Jialin, et al. "Stock Prices Prediction Using Deep Learning Models." ArXiv.org, Cornell University, 25 Sept. 2019, arxiv.org/abs/1909.12227v1.

[5] Matsunaga, Daiki, et al. "Exploring Graph Neural Networks for Stock Market Predictions with Rolling Window Analysis." ArXiv.org, Cornell University, 27 Nov. 2019, arxiv.org/abs/1909.10660v3.