Systematic Detection of Insider Trading

Report Link: https://github.com/11223548/UTS_ML2019_Main/blob/master/11223548_A3_ProposalFinal.pdf
Video Link: https://github.com/11223548/UTS_ML2019_Main/blob/master/11223548_A3_ProposalVideo.mp4
Course ID: 32513
Student ID: 11223548

Project Aims

The aim of this project is to develop a machine learning algorithm capable of detecting illegal insider trading in United States (**US**) stock markets. Construction of this algorithm involves successful completion of four important tasks:

(i) aggregation of tick-level trade and order book data for stocks listed on the NASDAQ, NYSE and AMEX;
(ii) preprocessing to convert data into insightful metrics that extract characteristics related to insider trading;
(iii) training, testing and evaluation of the performance of several machine learning approaches; and
(iv) selection and implementation of the optimal model for detection of insider trading

This proposal will outline why development of an insider trading detection algorithm is important; and how such a system would deliver significant regulatory and financial benefits to the US Securities and Exchange Commission (**SEC**) should they choose to invest in the project.

Background and Importance

Healthy stock markets are important for the efficient allocation of resources within modern economies. Since economic policy is often at the forefront of US politics, the prevention of activities that may cause significant harm to the functioning of the economy is of great interest to the US Government and its agencies. To this end, the US SEC is entrusted with the monitoring and deterrence of illegal insider trading in the US and should have a strong interest in investments that serve to accomplish these goals.

Is preventing insider trading beneficial to financial markets?

Academic research suggests that public markets must at least be perceived to be fair or else rational traders would not trade due to adverse selection and financial markets would collapse (see Milgrom and Stokey, 1982). Significant information asymmetry between investors would discourage participation in the stock market since less informed traders would have a very high chance of losing money on their investments. With less market participants, stocks would become less liquid, market pricing would become less efficient and stock valuations would decrease to reflect higher investment risks. Subsequently, resource allocation within the US economy would deteriorate; harming economic growth. This notion has been explored through empirical research. Investor appetite for equal access to material information is supported in Bhattacharya and Daouk (2002) who study 103 countries and show that enforcement of regulations that prohibit insider trading are rewarded by investors in the form of higher stock valuations across the market. Du & Wei (2004) also present evidence that links

greater enforceability of anti-insider trading regulations across countries with lower stock market volatility (a commonly used measure of market risk).

How have existing attempts to monitor and deter insider trading fared?

Equal access to information is clearly a concern to the US Government. As a result, the US Government and its agencies progressively introduced and enforced regulations aimed at preventing insider trading throughout the 20th century.[1] However, prohibition of illegal insider trading has been challenging to enforce due to difficulties in identifying illegal trading activities. Illegal insider trading involves the intentional obscuring of trading activities that seek to profit from material non-public information. As a result, traditional approaches to identifying instances of illegal insider trading have had limited success.

The ability to detect incidences of illegal insider trading in stock markets has remained elusive in financial literature. Published research on the observable features of illegal insider trading is contradictory. Cornell and Sirri (1992) and Fishe and Robe (2004) find conflicting results on whether bid-ask spreads are affected by insider trading. Ahern (2018) tests a variety of illiquidity measures and identifies that only absolute order imbalance and the negative autocorrelation of order flows are statistically and economically robust predictors of illegal informed trading. However, the results in the paper only hold for short-lived information. Ahern concludes that standard measures of illiquidity have limited applications for the detection of illegal informed trading. From literature it is unclear what observable market conditions are associated with periods of illegal informed trading.

Attempts to use measures of illiquidity to detect the presence of insider trading may be a fruitless endeavour. Since informed traders could be liquidity providers (limit orders), or liquidity takers (at-market orders), there should be no *ex ante* expectation that illiquidity measures would be able to identify the occurrence of informed trading in financial markets. Kaniel and Liu (2006) demonstrate that informed traders are more likely to place limit orders than market orders for long-lived information. Holden and Subrahmanyam (1992) and Back et al. (2000) show that where multiple risk-neutral traders have access to inside information competition induces immediate trading and therefore at-market orders are more likely.

From a regulation perspective, it may seem that illegal informed trading in financial markets should be *decreasing* over time due to the enforcement of increasingly stringent financial reporting requirements. Enhanced disclosure *should* reduce the opportunity for long-lived material private information that can be traded upon. By contrast, Acharya and Johnson (2010) finds that insider trading becomes more likely with more insiders in spite of increased regulation. Banerjee & Eckard (2001) analyse a wave of US mergers from 1897-1903, a time predating domestic legislation against insider trading, and find that pre-announcement stock run-ups and post-announcement price jumps closely resembled contemporary markets. The evidence is suggestive that illegal insider trading is still highly prevalent within the US.

Severe financial penalties may be insufficient to deter illegal insider trading in the US. Criminal prosecutions of illegal insider trading carry a maximum individual fine of US$5m and maximum prison sentence of 20 years. Fines for corporations can be much higher, with a record US$1.8 billion paid by SAC Capital in 2014 for their involvement in insider trading. The prevalence of illegal insider trading, despite hefty penalties, suggests that participants believe there is a low probability of being caught.

[1] Prominent 20th century US legislature restricting or deterring insider trading: Securities Act (1933), Securities Exchange Act (1934), The Insider Trading Sanctions Act (1984) and the Insider Trading and Securities Fraud Enforcement Act (1988).

An enhanced ability to detect participants in illegal insider trading would therefore provide a more significant deterrent.

In many cases the SEC has relied on whistleblowers to flag potential cases of illegal insider trading. These potential cases would subsequently be investigated further via interviews, examination of trading records and wire taps. Such a process typically involves extensive time and resources. Furthermore, a common reliance on third party reporting means the SEC will be unable to detect many incidences of insider trading where there were no whistleblowers. Evidently, there is a strong need for an alternate means to identify trading that has a high probability of representing illegal insider trading. The research project presented in this proposal has the potential to deliver a systematic process to focus SEC resources and deter illegal insider trading through increased detection rates.

Project Design and Innovation

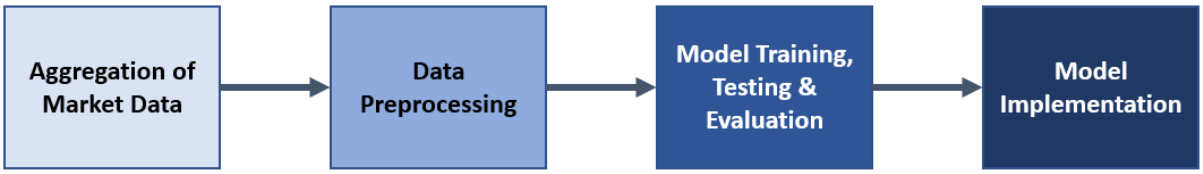
The key innovation of this research project is the application of recent developments in machine learning to the systematic identification of illegal insider trading. In particular, generative adversarial networks (**GAN**), which were only developed in 2014, represent a key component of the detection algorithm presented later in this proposal.

A shortcoming of prior attempts to detect illegal insider trading has been the reliance on linear models to identify characteristics representative of illegal insider trading. However, recent developments in machine learning represent a more plausible path to detection of illegal insider trading than traditional metrics explored in finance literature. Financial markets are typically complex, nonlinear and non-stationary in nature. Non-linear machine learning approaches are capable of capturing the complexity and non-linearity in such data. In a review of the recent use of machine learning in quantitative finance research, Emerson et al. (2019) conclude "*machine learning offers an opportunity for more complex financial analysis than was previously possible*".

Another distinction of this research project from prior attempts to systematically identify illegal insider trading is a focus on identification by exception. A common feature of – typically unfruitful - systematic approaches to identification of insider trading is an attempt to learn the features of trading windows in which a SEC-prosecuted case of insider trading occurred and then identify similar patterns elsewhere. Insider trading cases differ greatly in how the participants implemented trades and as a result the conventional identification approach will induce a lot of noise into the training of machine learning algorithms. By contrast, the algorithm in this proposal takes an opposite approach; it learns a representation of normal tick-level trading data and flags exceptions. This also makes the model less susceptible to manipulation since it does not observe prosecuted cases of insider trading until after the training phase.

How does this project approach the identification of illegal insider trading?

A high-level overview of the design of the research project is presented below.



Aggregation of market data

The first step of systematic detection of illegal insider trading involves the aggregation of tick-level order book data and trade data from the 3 major US stock exchanges (NASDAQ, NYSE & AMEX). Use of tick-level data is important since aggregation of trading windows could hinder the ability to identify unique trades and sequences that represent abnormalities. During model development, historical market data will be aggregated. However, upon implementation the final model would aggregate real time data which would be preprocessed and fed into the detection algorithm.

The detection algorithm presented in this proposal will initially be restricted to analysis of the top 500 largest listed companies across these exchanges. Illegal insider trading that affects these companies are likely to have the most detrimental impact on efficient resource allocation within the US economy. Furthermore, these companies will have sufficient availability of trading data to reliably train machine learning algorithms. Using both order book and trade data is also useful since it may be relatively easy for an inside trading to obscure the market impact of their illicit trades, so as to not be noticeable in trading data, but the anomalous order sequences placed to obscure trades would become easier to identify in the order book.

Data preprocessing

The second step of the detection algorithm would involve converting tick-level order book and trade data into a variety of metrics that could tease out the abnormal features of the time-series data. Whilst numerous metrics will be trialled, among the most interesting metrics are liquidity and aggression-based metrics. One example of an aggression metric would be price impulse; the price of a bid/(ask) order relative to the best bid/(ask) order in the market prior to the order's placement. A series of orders placed inside the spread would reveal one or more traders attempting to aggressively trade on information; prioritising execution time over best price.

Model training, testing and evaluation

A crucial component of the modelling underpinning the detection algorithm is the application of a standard modelling framework to each stock separately. The characteristics of each stock are unique as the valuations of each company are subject to idiosyncratic risks. As a result, any algorithm which attempts to aggregate characteristics across many different stocks will be less effective at determining what is an anomaly for a particular stock since it will be heavily impacted by noise.

In order to develop the detection algorithm, a GAN framework is utilised. However, a variety of alternate neural networks are to be trialled within the GAN framework. The default expectation is that a LSTM-GAN would be an appropriate model specification due to its ability to learn sequences of observations. Being able to analyse a sequence of observations should enable the detection algorithm to become more robust to noisy individual trades and instead flag sequences that are abnormal.

Each GAN is first trained on plausibly clean market data; one week of tick-level trade and order data more than three days preceding a company announcement. This enables the model to learn normal intraday trading patterns. Once trained, the discriminator is then fed data from the three days preceding a company announcement and makes predictions on whether each sequence is normal or abnormal based on its prior training. This second stage represents model testing.

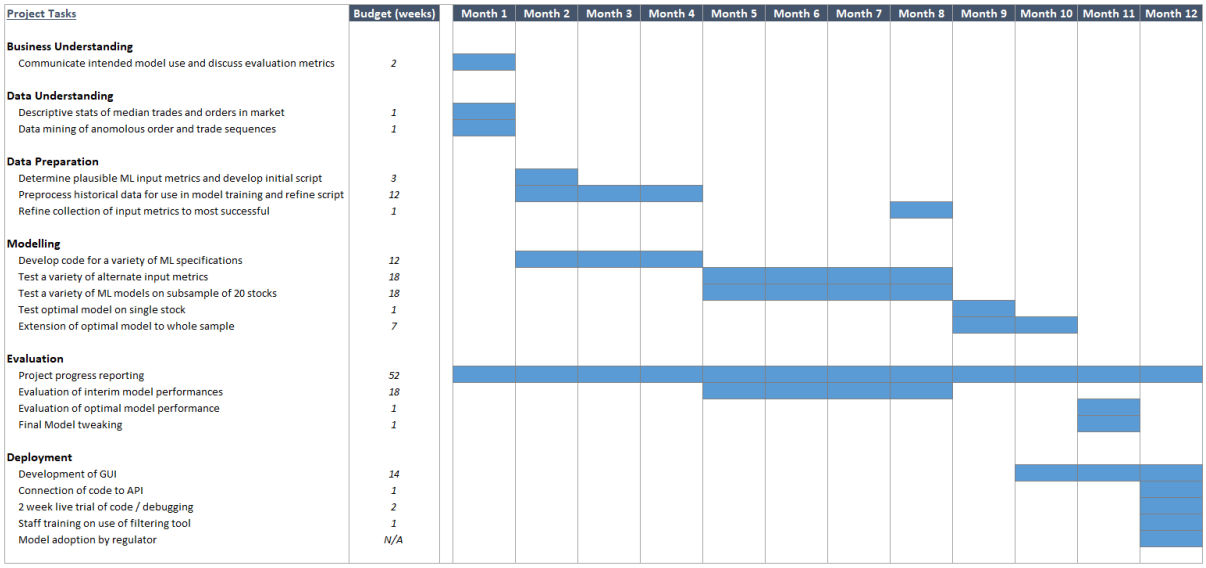
In evaluating the performance of various model specifications both precision and F1-scores will be considered. In this case the prediction of a trading sequence anomaly that aligns with a prosecuted case of insider trading would represent a true positive. The model framework which has the best performance across differing stocks will be selected as the optimal model.

Model implementation

Pursuant to the development of a successful detection model, a live version of the algorithm would be implemented on behalf of the SEC. A graphical user interface (**GUI**) would be developed by a software engineer which would make the algorithm easier for SEC staff to use and understand. The live algorithm would operate via a cloud-based computing cluster. Two weeks of formal training on how to use the algorithm would also be provided to SEC staff.

Project Development Timeline

Completion of the research project is expected to take 11 months. Following completion of research, the SEC will be required to decide whether they want to implement the detection algorithm developed during the research period. As part of their investment in this research project, the SEC is entitled to an early termination option at the end of the 11-month period which protects them against a significant ramp-up in costs from the implementation phase onwards if the algorithm is not producing satisfactory outcomes. The implementation phase would take one month and would include formal staff training on how to utilise the detection algorithm as a screening tool for insider trading investigations. A detailed timeline of project development is presented below.



Budget

	Development Costs (US$000's)	Ongoing Annual Expenses (US$000's)
Hardware		
HPC Cluster		
Net+ Google Cloud Platform - On Demand (10 x 32 cores) **(1)**	$21	$248
Personnel		
Data Scientist (*Myself*)		
Script development	$120	
Script maintenance & training		$10
Software Engineer 1		
Development of GUI for detection algorithm	$30	
Software Engineer 2		
Connection of algorithm to API, debugging	$10	
Total	**$181**	**$258**

Notes:

1. Cost is US$0.024815 per core hour. Assumes 12 hours of use per day, 5 days per week (reasonable given market is open for ~7 hours per day on weekdays only).

2. Incentive payments are excluded from the budget since the incidence of insider trading prosecutions and extent of fines revenue is highly variable.

The project budget has been structured in a way that mitigates risks and upfront costs to the investor. The bulk of development costs are incurred by the SEC only after proof-of-concept has been established. If the detection algorithm does not produce sufficiently accurate results then the SEC can exercise their early termination option. Personnel costs are in line with typical market salaries for the region and have been benchmarked in the Personnel section of this proposal.

Development costs

Development costs budgeted for the project are unusually low for a project that can deliver such a large amount of potential value to the SEC. This is because the project has been designed to mitigate investment risks for the SEC and instead provide entrepreneurial incentivisation through performance-linked incentive costs. The initial development costs relate to 12 months of salary for a data scientist, 3 months of salary for one software engineer, 1 month of salary for another software engineer, and 1 month of access to a cloud-based computing cluster. Through exercise of their early termination option the SEC could avoid one month of personnel costs and cluster costs.

Ongoing costs

The primary fixed ongoing cost relates to operating a computing cluster to run the insider trading detection algorithm. Operating the algorithm via a third-party service provider was found to be far more cost-effective than internally building a computing cluster since third parties such as Amazon and Google have significant economies of scale. Of the available reputable service providers, Google was identified as the cheapest option and also offered important security features. An on-demand service was more cost-effective than a continuous service since the detection algorithm is only required to operate on market days and around market hours.

In addition, there is a small recurring data scientist cost budgeted for maintenance of the detection algorithm and training of SEC staff.

Incentive costs

The SEC's investment in this research project has been structured to minimise their risk whilst ensuring that development of a high performing detection algorithm is strongly incentivised. In order to align interests, the SEC would agree that, for the first ten years of algorithm implementation, 1% of future fine revenue generated from insider trading prosecutions would be payable to myself as an incentive fee. This ensures that I am highly motivated to develop a detection algorithm that is capable of identifying insider trading. By comparison, the data scientist salary during the development phase reflects standard market rates and would be insufficient to entice entrepreneurial development of a high performing detection algorithm.

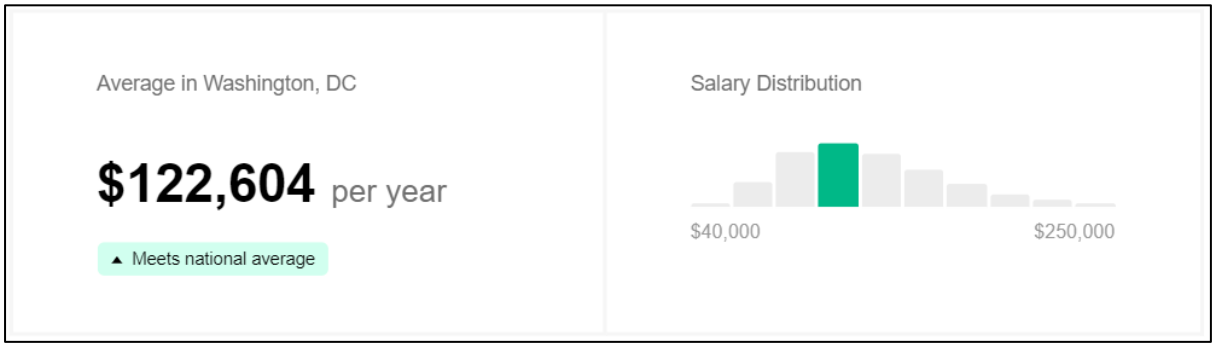
Personnel

It is assumed that the development of the insider trading algorithm will occur on-site at the SEC's headquarters in Washington, DC. Working on-site will facilitate collaboration and progress reporting with the SEC throughout the project. To reflect the reasonableness of the staff salaries during the project, each salary was benchmarked against salaries for similar roles in Washington, DC.[2]

Data Scientist

[2] The US version of indeed.com was accessed to benchmark salaries. Salaries were filtered to represent Washington, DC which is the where the US SEC headquarters are located.

Throughout the 12-month project duration my primary role will resemble that of a data scientist. I will be the sole person responsible for developing the detection algorithm and testing alternate models. Once the algorithm has been implemented, I will also be responsible for training of the SEC staff on how to utilise the detection algorithm as a screening tool for identification of trades that most likely represent illegal insider trading. For this role I have allocated an annual salary of US$120,000 to be paid monthly, in advance. As illustrated below, this cost is in-line with market salaries for a data scientist located in Washington, DC. After the initial 12 months of development I have budgeted one month of time per year of ongoing costs to review the performance of the detection algorithm and potential enhancements as well as further staff training.



Source: Indeed.com

Software Engineers

Two software engineers will be required during the development of the detection algorithm. In the final three months of the project one software engineer will be required to develop a graphical user interface (**GUI**) that will make the detection algorithm easier for SEC staff to use and understand. In the final month of the project, an additional software engineer will be required to focus on connection of the detection algorithm to live data feeds and implementation of the algorithm on the Net+ Google Cloud Platform. The cost budgeted per software engineer is US$120,000 per annum pro-rata to the employment duration. As illustrated below, these costs are slightly above market salaries for software engineers located in Washington, DC. The higher costs are due to the employment period being relatively short-term and therefore requiring greater financial incentive to attract software engineers.

Source: Indeed.com

Summary of Key Benefits to the SEC

This research project has the potential to deliver significant benefits to the SEC whilst also mitigating investment risks. A robust insider trading detection algorithm would enable the SEC to flag a much larger variety of illegal trades and result in a higher prosecution rate. This would strongly deter financial market participants from engaging in illicit trading since they would recognise there is a much

higher chance of getting caught. In turn, this would build greater public confidence in the US stock exchanges and induce higher participation in the stock market; benefiting the resource allocation within the US economy. A systematic approach to identification of insider trading would also enhance resource allocation within the SEC. The time of staff could be focused on investigating the most likely cases of insider trading.

The research project proposed has been structured in a way that significantly mitigates investment risks for the SEC. Costs during the development phase of the project are limited, with many of the costs being backended or conditional on performance. The SEC would hold an early termination option on the project which would prevent the accumulation of further investment costs if the project did not show significant promise at the end of the development phase. In addition, the remuneration for development of the detection algorithm is heavily dependent on the actual performance of the algorithm due to the incentive fee structure proposed. This greatly reduces any incentive to present misleading results.

Altogether, this research proposal represents an excellent, low-risk investment opportunity for the SEC to improve their enforcement capabilities.

References

[1] V. Acharya and T. Johnson, “More Insiders, More Insider Trading: Evidence from Private-Equity Buyouts”, *Journal of Financial Economics*, vol. 98, no. 3, pp. 500-523, 2010.

[2] K. Ahern, “Do Proxies for Informed Trading Measure Informed Trading? Evidence from Illegal Insider Trades”, *Working Paper*, 2018.

[3] K. Back, C. Cao and G. Willard, “Imperfect Competition Among Informed traders”, *The Journal of Finance*, vol. 55, no. 5, pp. 2117-2155, 2000.

[4] A. Banerjee and W. Eckard, “Why Regulate Insider Trading? Evidence from the First Great Merger Wave (1897-1903)”, *American Economic Review*, vol. 91, no. 5, pp. 1329-1349, 2001.

[5] U. Bhattacharya and H. Daouk, “The World Price of Insider Trading”, *The Journal of Finance*, vol. 57, pp. 75-108, 2002.

[6] A. Bris, “Do Insider Trading Laws Work?”, *European Financial Management*, vol. 11, no. 3, pp. 267-312, 2005.

[7] S. Emerson, R. Kennedy, L. O’Shea, and J. O’Brien, “Trends and Applications of Machine Learning in Quantitative Finance”, *8th International Conference on Economics and Finance Research*, 30 May, 2019.

[8] R. Fishe and M. Robe, “The Impact of Illegal Insider Trading in Dealer and Specialist Markets: Evidence from a Natural Experiment”, *Journal of Financial Economics*, vol. 71, pp. 461-488, 2004.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets”, *Neural Information Processing Systems, 2014*.

[10] D. Guercio, E. Odders-White and M. Ready, “The Deterrence Effect of SEC Enforcement Intensity on Illegal Insider Trading”, *Working Paper*, 2013.

[11] C. Holden and A. Subrahmanyam, “Long-Lived Private Information and Imperfect Competition”, The Journal of Finance, vol. 47, no. 1, pp. 247-270, 1992.

[12] A. Jeanes, “Cloud vs. Datacenter Costs for High Performance Computing (HPC): A Real World Example”, *Internet2 Blog*, weblog, 19 June 2017, <https://www.internet2.edu/blogs/detail/14114>

[13] R. Kaniel and H. Liu, “So What Orders do Informed Traders Use?”, *The Journal of Business*, vol. 79, no. 4, pp. 1867-1913, 2006.

[14] P. Milgrom and N. Stokey, “Information, Trade and Common Knowledge”, *Journal of Economic Theory*, vol. 26, no. 1, pp. 17-27, 1982.