Placement Report

Jihye Hong

Submitted for the Degree of Master of Science in

MSc Data Science and Analytics

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

1. About the company

The company 'Lindgren Laboratory' is a financial technology company based in London. The company was originally founded in 2010 in Gothenburg. The business is one of the world's leading companies for pricing of Andrew Debreu contracts. Its customers are mainly from the financial and sports betting industries. The company has grown with the development of automation with statistical modelling in the industry and now have a massive interest in new approaches driven by machine learning and artificial intelligence. The business aims to develop and deploy the best statistical pricing model in the industry. One of the current main research areas is the improvement of existing predictive models in sports betting industries. The company is currently working on football, tennis and cricket, but also looking to expand with other opportunities in other sports betting markets. Moreover, another main focus of the business is the high-frequency market making. The company is taking it further from their successful traditional business models to new areas of emerging financial markets such as the cryptocurrency markets.

2. The role

As a quantitative analyst in the company, the ultimate focus of my role was the development of rigorous algorithmic trading and pricing models in order to achieve excess returns in various financial markets. The main task of the role was on the iterative research-driven environment for the development of algorithmic trading software on behalf of clients. The work is highly independent. It allowed me to experience various areas of data analysis projects from scratch. The role required various skill-sets under different tasks. Regarding data analysis and visualisation, at first, it was easy to adapt myself to use R in practice, due to it was the main tool that I exploited for the MSc courses. Also, C# was the main language in the company to develop software programs. The major language I have used in my previous job was Java. However, Java and C# have similar features as object-oriented programming languages. From that perspective, it was not that hard to learn C#. For database management, one of the popular relational database MySQL was on the top of the other development stacks.

3. Aims and objectives

3.1 Personal Aims

Concerning my personal goals for this placement, it was applying theoretical knowledge learnt from the course to practice in real-world data. When the placement started, it was a little bit confusing which area of data science career would suit me. My first goal in this degree was learning about big data processing to be a data engineer. As my previous work experience was from software engineering in Java, large-scale data processing with Hadoop or MapReduce was one of my keen interests with a lower barrier than statistical subjects. However, once the course started, machine learning and data analysis subjects attracted my interests from me. My ultimate goal is to be a leader of a data science team in combining my programming experience with data science. For that goal, it is

essential to know both data engineering as well as data science with statistical aspects. Therefore, which skill-sets would be developed from this placement job was highly important to me for the first stepping stone. From this perspective, my aim from this placement role was to fulfil the lack of knowledge and experience in statistical modelling and apply machine learning methods in practice. Also, as carrying out various data-driven projects, it allowed me to have hands-on experience and in-depth knowledge of new technologies such as Python, PySpark and Kafka.

3.2 Business Aims

The business aim from this placement position was to research and develop a new business model via high-frequency trading strategy in the cryptocurrency market. As the cryptocurrency market is emerging and highly volatile, the company believes it holds a wide range of new trading opportunities. In particular, applying various benchmark models from traditional financial markets such as the stock market was one of the main focus. So, the business required me to quickly cycle the process of design, develop and deploy proof of concept programs to prove if the given trading model show the evidence of excess returns.

3.3 Day-to-day Work

There are two main areas that I have worked on. One is the development of novel trading strategy in the cryptocurrency market, the other is research to improve the existing football forecasting model. Cryptocurrency project was 100% independently carried out on my own apart from guidance and suggestion about the strategies from my manager, and senior modellers in Sweden. During the placement, I have achieved to gain problem-solving skills for data-driven projects with improved data analysis and machine learning skills in business. Also, the experience provided a wide range of hands-on experience in practical data analysis projects.

To name a few of day to day jobs, it includes:

- Raw data collecting and processing from public sources, cleansing dirty raw data into meaningful information, extracting them into databases.
- Exploratory Data Analysis and statistical modelling in financial markets using SQL for data manipulation, using Python for prediction of market movements in real-time with machine learning binary classifiers.
- Automation of algorithmic trading programs, design system architecture, model databases, develop and execute prototypes in financial markets. Programming in Java and Python.
- Built comprehensive data streaming pipeline with Kafka for integration of data sources.
- Developed a dashboard as a Spring MVC Java Web Application with D3.js, demonstrated summary of trading, underlying trends in the market data.
- Work with Git for source control and experience in shell scripting in Linux environment.

4. Projects

There have been a number of small projects done in different focusing areas, this chapter will demonstrate the aim and description of projects undertaken.

4.1 Statistical arbitrage trading

The goal of this project was to address the feasibility of continuous high-frequency trading with the underlying idea of statistical arbitrage trading. As statistical arbitrage is a widely used algorithmic trading model, this project aimed to show how the benchmark will perform in the cryptocurrency market. A distinct idea of this project from the traditional way was to introduce 'triangular' and 'inter-exchange' arbitrage trading. Arbitrage trading strategies make a profit from the inefficiency of prices between securities. The algorithmic-trading involves simultaneous buying and selling of securities in a conventional way. The clear idea of this project was putting the buy and sell orders in a triangular shape as trading three different cryptocurrency pairs in a row in one exchange named 'triangular arbitrage'. Moreover, a strategy called 'inter-exchange arbitrage' to put one buy order in the exchange A and one sell order in the exchange B to exploit the inefficiency of different exchanges.

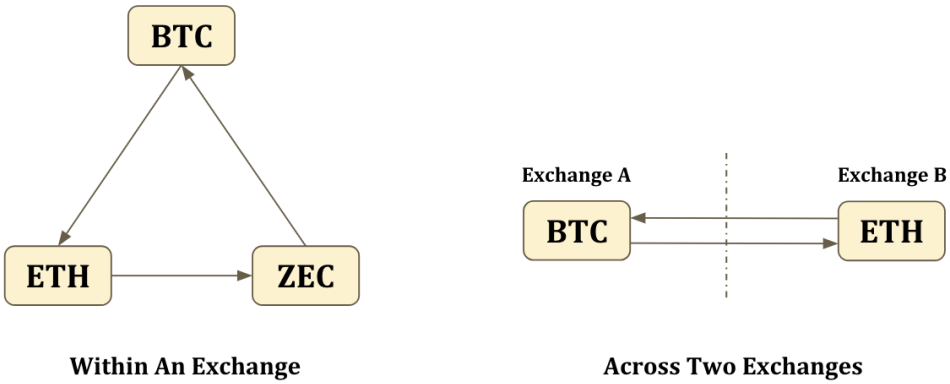

Figure 1. Arbitrage models

The project allowed me a number of new opportunities. First, I have learnt C# to develop algorithmic trading bots. As C# is the main development stack in the company, I was suggested to develop programs in C# at that stage. Second, I designed and maintained MySQL server to store trading logs and historical price data from trading agents. In the meantime, ETL process from raw data collected via REST APIs was written in SQL on the top of C# to manage the data connection and manipulation from the agent side. Moreover, I designed the physical and logical relation of database objects while I was learning the domain knowledge of the financial industry, including the basic trading terms such as 'order book', 'bid-ask spread' with well-known research studies like the Turtle trading system.

Overall, it was easy to adapt to the development stack of this first project with the previous working experience as a Java software developer. Although Java and C# are two different languages, their nature and features are somewhat similar. Therefore, the initial version of the proof of concept program for arbitrage trading had been developed quicker than my manager expected. However, after deploying the program to run in practice, several problems have arisen such as low-latency and high-frequency design of software. However, as I encounter those problems, it improved my software engineering skills I was able to learn how to manage multi-threading in high-frequency trading models to develop better-performing agents.

Although this initial project was successful concerning continuous trading and high turnover with high-frequency, it was not the best timing to test the new trading approach. The cryptocurrency market peaked in December last year (2017), then it almost crashed. As a result, the market was too volatile and hard to evaluate the actual profit earned from the trading model. However, as a first stepping stone to build a robust algorithmic trading model, we confirmed its rising trading volume and market capitalism. Hence, it was continued to find a better trading opportunity finding insights from the collected trading logs and price data.

4.2 Dashboard development with D3.js

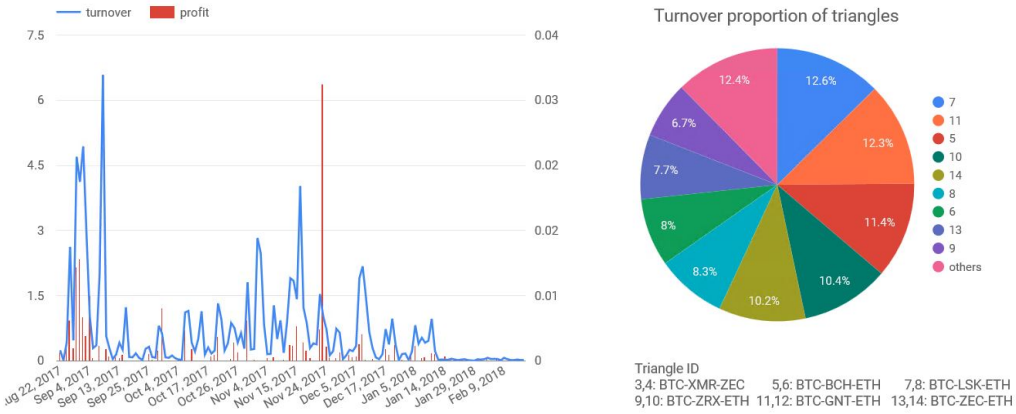

Figure 2. Plots in the dashboard

The trading logs collected from real-time trading data in the crypto exchanges were highly valuable to build a business intelligence dashboard to demonstrate the KPI and performance of the arbitrage trading agents to stakeholders. To design and develop the dashboard, a web application in Java was developed. To get useful insights from the data, I developed that the application connects to the MySQL server via SQL queries. The front-end side was developed with Javascript, jQuery and HTML. Also, D3.js was used to plot the data using web-standard. D3.js is an interactive data visualisation library in javascript. It was a great opportunity to have a hands-on experience with D3.js. Nowadays, D3.js is one of the highly desirable skill-set to be an expert in data visualisation. Recently, there are many companies working with D3.js especially for real-time dashboard design. Therefore, I have chosen to use D3.js to demonstrate the essential information of trading performance to stakeholders. It was genuinely fun to explore the interactive

data visualisations. It was a great experience to know the advantages and disadvantages of interactive web visualisation over static plots drawn by R or Matlab.

The dashboards consist of several parts. It demonstrated the daily turnover and number of trades with the details in history. It was fun to learn d3.js for the first time as it has various useful functions and features to deploy an interactive web dashboard. Also, I could exploit what I learnt from the Data Visualisation course regarding how to design better plots for clear interpretation as well as developing my exploratory analysis skills.

4.3 Data Analysis for comparison of exchanges

The next project was data analysis for cryptocurrency exchange comparison. The underlying idea of this project was there is the true price in the market. Therefore, through data analysis, we aimed to address which exchange is efficiently following the true price and which are not. The underlying assumption came from the 'cross(inter)-exchange' arbitrage trading in the first project as we explored two cryptocurrency exchange to research it. With this theory, when two exchanges have the distinct price difference, the market should adjust its inefficiency. But from the initial research result for inter-exchange arbitrage trading, it seemed there are certain trends of lower or higher pricing in exchanges.

To explore the area of interests, five different exchanges were selected to be compared. There are hundreds of crypto-exchanges nowadays. Some only deals with crypto-crypto markets, while some markets have crypto-fiat markets to sell or buy the coins in fiat money such as US Dollar or British Pounds. The figure below depicts the historical price of five exchanges. It shows that 'Kraken' has a trend of pricing lower than others. From this analytical view, we applied the inter-exchange arbitrage strategy in the two exchanges 'Bittrex' and 'Kraken' which have the highest and lowest price trends. Previously the strategy has targeted 'Bittrex' and 'Poloniex'. However, there were not many arbitrage chances with a small average daily turnover. As we applied the same strategy towards 'Bittrex' and 'Kraken', it tripled its turnover value. It resulted in proving the bigger inefficiency across different exchanges would return excess profits in arbitrage trading. It was a significant finding to move forward in this given problem.

While I carried out this project, it brought some fundamental questions about data analysis and hypothesis test. For example, we developed the trading model on randomly chosen two exchange on the first project. The exchange comparison analysis uncovered that those two exchanges have very similar prices over time. Consequently, the model was not highly profitable.

Like the example, diving into develop, deploy models first followed by analysis afterwards would consume a considerable amount of time with a potential waste of time when the designed model has a dead-end. On the other hand, the real application of a model more or less always likely to show unexpected results which are different from the analytical result. This trade-off gave some insights on how to plan timelines for designing a research project. It was a useful understanding through the placement projects. Also, I could understand the efficient design of data analysis projects and developing models.

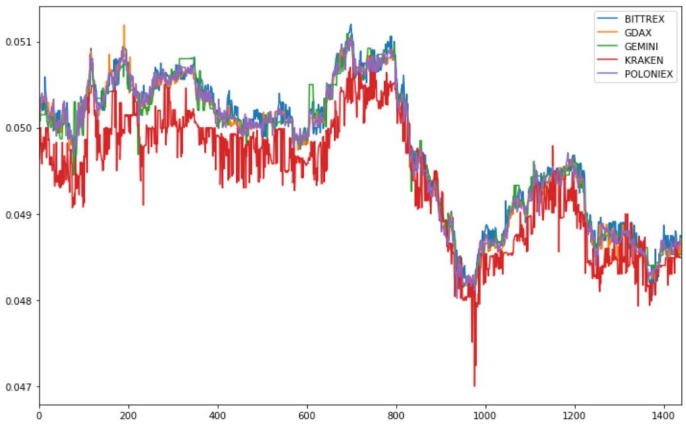

Figure 3. Historical Ethereum price in exchanges

4.4 Market Making trading bot on EtherDelta

This project aimed to explore the trading opportunity on EtherDelta, which is a P2P decentralised cryptocurrency exchange using Ethereum blockchain network and technologies. EtherDelta has smaller trading volumes than major exchanges but it has large bid-ask spreads than others. Thus, we decided to investigate if a simple market making algorithmic trading model would work on this new and different type of exchange. I have learnt low-level of blockchain technologies through this project such as cryptography and smart contract. Also, the poor documentation of APIs of EtherDelta made me spend a huge amount of time to build a trading program from scratch with a deep understanding of the fundamental technology of Ethereum. However, when I was working on this project, there was a blockchain-based game called 'CryptoKitties' was introduced by a developer. CryptoKitties completely occupied the resource of Ethereum network. Consequently, it was tough to confirm a buy or sell order on the network without consuming huge 'gas', which is a kind of commission paid to execute a transaction on the network.

Moreover, the exchange was hacked and sold to a new owner after a few months of the event. Since then, the exchange is not functioning correctly. As a result, this project was finished without any success.

4.5 Trading with Machine Learning models

This project was based on a number of approaches to predict the price movements in the cryptocurrency market. This was designed to investigate the forecasting power of the historical price data. With exploratory analysis first, I designed several unique experiments to demonstrate the predictability of price movements. During this stage, many efforts were spent to research past literature and application in industries. From several benchmarks, I initially tested more than 10 different classifiers with a range of parameter setting. It was a great chance to exploit what I have earnt from Machine Learning and Data Analysis course in the degree towards real-world data. At the end of this project, we were able to show the earning from the machine learning approach is more profitable than a naïve random

model. However, the most effective time unit for the forecasting was an hour from the result of the project. Although this project draws attention to address how machine learning classifiers perform in the cryptocurrency market, it was hard to use the model for high-frequency trading. The strategy is rather suitable for long-term trading.

Nevertheless, it allowed me to learn many things and eye-opening experience as this project gave me an opportunity to widen my knowledge in online machine learning. Also, I grasped the research idea for predictability of price movements in financial markets. It has been explored as my individual project and thesis.

In this project, the development stage of the proof of concept program was divided into two phases. First is the development of a trading agent, and the other is the forecasting model with real-time market data. To carry on the given requirements, I decided to develop the program in Python. It was the first project I fully conducted in Python. As I thought, it would bring me great benefits with the various features for data science stack, also, as a functional programming language. Although the Python language was not originally designed for statistical computing like R or Matlab, with a wide range of statistical and machine learning libraries, the function as a statistical computing language is the greatest benefit of Python. I could have opportunities to explore the full range of Python's Data Science stack through this short project to establish the first stepping stone in learning Python.

Furthermore, I developed data pipelines to integrate the price data part and trading data part from the market with Apache Kafka, which is a low latency platform to build real-time data pipelines and streaming.

4.6 Football Prediction Models

The next project I took part in aimed to develop a new generation of a football prediction model. One of the namely well-known products of the business is predictive models for the clients in the sports betting market. The existing model was based on traditional statistical methods 'Monte Carlo'. Although when the model was launched in the market for the first it dominantly outperformed the conventional models, after several years, the performance was a little bit behind within the market. In order to examine how we can build a new model with a great performance, several ideas have been previously explored such as Logistic regression, Poisson distribution and Maximum Log-likelihood internally by senior modellers. However, there was not much luck in the research, as they did not outperform remarkably over the current Monte Carlo model. That was why the company hired data science consultants to see the given problem out of the box. The company hired a data science consulting firm 'ASI Data Science' to carry out this project. It was a great chance to learn how the data science team accomplish the requirements of a data science project from scratch. It was quite impressive the way how ASI specify the problem quickly and precisely as well as demonstrate the outcomes in detail. The ASI data scientist used Python and I was the only member of my company who has experience in Python. So, when it was necessary, I interpreted the code in R. It was the first project I participated in with other modellers in the company.

To work with the team, the main tool used was 'Sherlock ML' which is a centralised data science terminal developed as a cloud service of 'Jupyter Notebook'. The fact that I carried on the last project in Python helped me to follow the work although I was still not 100% familiar with Python Data Science stacks at that moment. The codes the consultants developed were useful resources to benchmark their progress. My task was testing a simple Deep Learning LSTM model for the binary classification of football match results. I used one of the most popular deep learning libraries, Python Keras. Also, I self-learnt by myself the underlying concepts of Deep Learning with the lecturers' note of 'Deep Learning' course on Moodle.

As I carried out this data science project, I learnt how to evaluate the model performance under the given criterion with precisely specified metrics. In the end, the newly suggested models from the project were not very successful as compared to the current model. The fact is it is rather hard to evaluate the performance of the model in practice, as there are so many factors that can affect to the model by each game. As a result, this project is still in the area of development to find a novel approach to improve the prediction performance.

4.7 The sentimental analysis in the cryptocurrency market

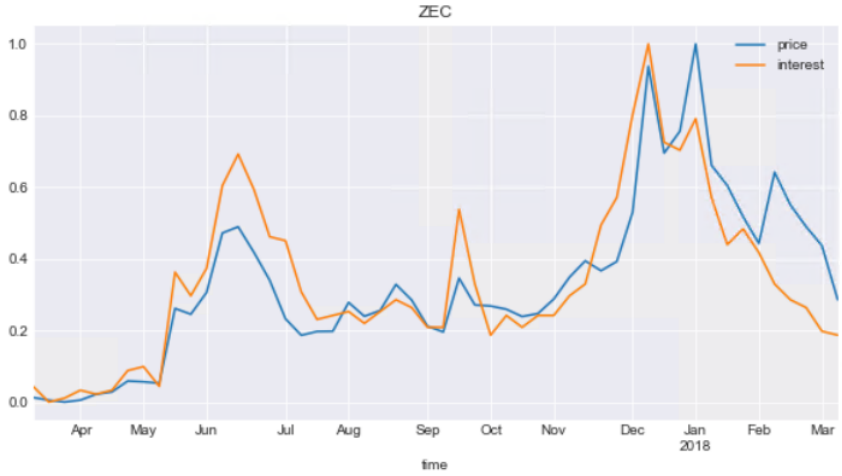

Figure 4. Google trends vs price of Z-Cash

The fundamental idea of this project is to show the correlation between Google search interest and the price of cryptocurrency. There is a service called 'Google Trend' provided by Google. It is a public web service to gain insights into search keywords in terms of its trend, popularity, etc. It can be a useful tool to understand and explore what is happening with the underlying social phenomenon with the given keywords. The main goal of this project was to reveal if the price of cryptocurrency is responsive or predictive according to the Google Trends data. As a result, search trends of major coins such as Bitcoin or Ethereum have so much noise and looks quite responsive to the price. Google Trends data showed when the price soared or plunged the interests in the keyword also highly increased. On the contrary, Google trends of small alt-coins like 'ZCash' or 'Monero' showed some

predictive power to a certain degree. As the stakeholders prefer to explore more traditional ways of trading strategies to a quick outcome, there was no further chance to explore the research area. But through this project, I had a massive interest in natural language processing and I could learn the basic concept of natural language processing. Moreover, it was a great chance to explore other Google products in data such as Google Data Studio and BigQuery to build a simple dashboard or an interactive web report.

4.8 Market making agents in the cryptocurrency market

In this project, I was in charge of the development of automated algorithmic trading agents with traditional market making strategies. One of the interesting topics in this project was the volatility analysis with real-time market price data. Therefore, I managed to develop an agent with a flexible margin which is adaptable to the market volatility. For example, when the market has very small bid-ask spread and there is a very small chance of trading for the agent program with the default margin rate. To resolve the problem, the agent automatically traces the current market volatility and adapt to the level of volatility by adjusting the margin rate.

During the progress of this project, it required me to have a number of discussions about financial trading strategy with my manager in detail. As my background is not from finance, I have had a lack of knowledge in the financial domain. While I have worked on this project, it gave me a plentiful of opportunity to learn the underlying concept of market making with relevant economics, such as risk management and assessment of profit & loss. I am currently working on this project as well and developing many areas related to the given problem. I believe this experience will be an absolute help to broaden my horizon to pursue a career in financial technology industries.

5. Overall experience

5.1 Employment Process

For the employment process of this company, I had 3 interview stages, telephone interview, face-to-face interview, and the technical test. Through the process, I could have understanding in job application process in the UK and demand of skill-sets in data science industries.

5.2 Working Environment and Culture

The company provides a very flexible working environment. I could easily manage to work from home once a week or twice a week if it was necessary. Also, the position was highly independent and flexible without a fixed deadline. As the nature of work is research-based, it required me to keep focusing on developing a new area of business model or new hypothesis to test in practice. Continuous finding research topics and novel approaches to test out in quick cycles were somewhat unfamiliar to me, due to my background are from working in production environments of software development. Many dead ends by research made me a little bit exhausted sometimes. However, a couple of months later, I could manage to get used to the nature of the work. This flexible environment made me more

productive from time to time. In the ends, I could achieve skills to find appropriate resources to develop new projects and execute the concept from benchmark models. Moreover, the monthly review was held by my manager and I could see my achievements via that meetings.

5.3 Team

In the London office, I was the only full-time employee apart from my manager. As there is no senior developer or analyst in the office, it was rather hard to get along with the team in Sweden and get feedbacks at the right times. Although the team used co-working tools such as 'Slack', it was not completely as quickly as face-to-face meetings. In last October, there was a chance to visit Gothenburg to demonstrate the progress of statistical arbitrage trading in the cryptocurrency market. The communication process was much efficient and quick than online-communication.

From the experience, I realised again that I really enjoy working in a team with regular feedbacks from colleagues to appreciate my work as well as learn from each other. If I were a student without previous working experience, it would have been very tough to get used to the independent work environment.

5.4 What I felt from this experience

Apart from what I learnt in the work projects, in general, it was full of opportunity to improve my problem-solving skills. In the meantime, the work environment allowed me to adapt to the decision making under the start-up environment. It made me achieve great adaptability in any kind of organisations. Overall, this placement year has offered a range of interesting work with good people. It was such an eye-opening experience to learn cutting-edge technologies in a whole different industry from my background. Moreover, I could learn how quickly growing the data science field concerning new technologies day-to-day. The majority of popular tools in this era are released in the past 10 years. For example, Apache Spark is now the notion of data engineering in industry, although it was released only 4 years ago. In particular, with the advent of open source products, it is vital to keep up-to-date with new technology. As well as, we should have standards for what technologies to learn or stay. As a quick learner, I am pleased to be a part of this rapidly developing industry. From this placement experience, I look forward to developing my career as a data scientist in the financial field.