**Advanced Technology Attachment Programme (ATAP)**

**Final Project Report**

**at**

**Lynx Analytics**

**Reporting Period:**

**01 2021 to 06 2021**

by

Tobias Yap

Department of Computer Science

School of Computing

National University of Singapore

2020/2021

Project Title: Data Science Intern (STC - Hwawei CEI+)

Project ID: AY20165656

Project Supervisor: Prof Heng Cheng Suang

## Summary

The following report summarizes my six-month internship experience with Lynx Analytics as a Data Science Intern. Through my time in the company, I went through a basic onboarding training for the first month doing tutorials related to graph theory on LynxKite, a proprietary software for graph visualization. For the following months, I worked together with several data scientists on the Saudi Telecom Company (STC) – Huawei CEI+ project. The project was in the first phase of Data Discovery & Understanding of existing fields with data, where I completed data understanding assignments regarding customer network experience variables using PySpark. Towards the end of the internship, the project moved into the next phase which was producing modelling outcomes and final ingredient selection for the Consumer Happiness Index (CHI) model to be implemented. The details of work done, and personal experience have been presented in the report below.

Keywords:

Data Science, Python, PySpark, SparkSQL

## Acknowledgement

I would like to express my deepest appreciation to my project supervisor, Senior Data Scientist Dr. Sergii Zakharov, Makara Phav and Swan Htet Lin whom I worked with on the STC project for sharing their extensive knowledge and valuable advice in the field of data science and for their continued guidance throughout my internship at Lynx Analytics. I would also like to extend my sincere thanks to Tharinda Premasiri, CIO Dr. Gabor Benedek and the rest of the members of the STC project whom I have worked with for the help and support they have provided me with throughout the project. I am also grateful to Director of Data Science, Marton Szel who supervised me during my stint on the Imagine project. Finally, I am deeply indebted to CEO Gyorgy Lajtai and Lynx Analytics for providing me with the opportunity to work on high profile projects, where I have come to learn so many new things regarding business consulting, big data and data science.

Table of Contents

[Summary 2](#_Toc75396035)

[Acknowledgement 3](#_Toc75396036)

[1 Introduction 5](#_Toc75396037)

[1.1 Background and Organisational Structure of Host Organisation 5](#_Toc75396038)

[1.2 Position of Host Unit within Host Organisation 5](#_Toc75396039)

[2 Training Assignments 6](#_Toc75396040)

[2.1 Training Assignments Completed in 1st Month 6](#_Toc75396041)

[3 Knowledge and Experience Gained 7](#_Toc75396042)

[3.1 Technical Knowledge Gained from Assignments 7](#_Toc75396043)

[3.2 Organisational/Industry Experience Gained from Assignments 8](#_Toc75396044)

[4 Conclusions 9](#_Toc75396045)

[4.1 Summary of Work Completed and Training Received 9](#_Toc75396046)

[4.2 Personal Experience 9](#_Toc75396047)

## 1 Introduction

### Background and Organisational Structure of Host Organisation

Founded in 2010 and headquartered in Singapore with an engineering team based in Hungary, Lynx Analytics provides strong expertise in predictive analytic models to help businesses in Telecom, Banking, Retail and Pharmaceutical fields enhance customer experiences and predict outcomes such as Churn rate, Customer Lifetime Value, Net-Promoter Score (NPS) and Average Revenue Per User (ARPU).

Lynx works with many companies internationally and typically deploys dedicated teams on-site, but due to the recent COVID-19 situation much of the work has been done remotely.

Lynx Analytics follows a functional organisational structure, departmentalizing the organisation based on common job functions like data scientists, data engineers, business consultants, marketing, finance etc. With CEO Gyorgy Lajtai at the helm, employees are assigned projects with different companies, some working on multiple projects in parallel.

### Position of Host Unit within Host Organisation

I was part of the data science team at Lynx Analytics, and more specifically, tasked to the STC project team.

## Training Assignments

There was an onboarding training for new interns that lasted around a month.

### Training Assignments Completed in 1st Month

Lynxkite tutorials completed and concepts learnt:

1. Graph theory for LynxKite users
   * Graph types, vertices, edges
   * Degree
   * Segmentation/clustering
   * PageRank
2. Exploring LynxKite
   * Platform navigation, workspace creation
   * Using platform features and filters to visualise graph attributes
3. Exploring a Facebook graph
   * Importing workspace
   * Visualizing 3D/bucketed graphs
   * Using graph metrics degree, PageRank and dispersion
   * Segmentation and communities
4. Analysing Airports and Routes database
   * Importing vertices and edges tables to workspace
   * Merging parallel edges
   * Modular clustering
   * Aggregating to/from segmentations
   * Deriving vertex attributes
5. App Store Fraud Detection
   * Aggregating to vertices
   * Merging vertex attributes
   * Utilising reports of fraudulent activity and finding out apps with common connections to flagged app developer for further investigation
6. TextRank
   * Reproducing TextRank and Levenshtein distance in LynxKite
7. Wikivote
   * Finding the number of triads on Wikipedia vote network
8. Rakuten – On-Demand Videos and Users
   * Create co-watching graph of videos
   * Regression model to identify relationship between episode count and number of distinct users watching the videos
   * Modularity and meaningful clustering of users by video watched and co-location
   * Creating tripartite graph
   * Find which two users watched the same videos at the same time consistently
9. Co-Location Analysis on Mobile Ad Data
   * Co-location: connect two vertices if they are at the same place, at the same time
   * Calculating number of distinct days two people are co-located to establish relationship and building communities

Much of the LynxKite material was informational and interesting, though the program itself was not used for the remainder of my time at Lynx I would be interested in furthering my knowledge in the field of Graph Theory.

## Knowledge and Experience Gained

### Technical Knowledge Gained from Assignments

**Setting up a workspace (VPN)**

Since I was working for Saudi Telecom Company (STC), I had to set up a remote workspace to be able to connect to their VPN. To do so while still having access to the internet, we had to use a virtual machine (Oracle VM VirtualBox), and for tunnelling we use Ubuntu. With the proper ssh config and VPN connection, we are able to access files in the HDFS within the terminal.

**PySpark, SparkSQL and data science**

Learnt and utilized spark code templates and functions to be able to work with big data. As part of data ingestion and understanding of STC’s customer data, I was made to work on deriving distributions of customer network experience variables. This involved a mix of creating spark sessions and running data aggregations with SparkSQL and exporting them to the HDFS for future outcome analysis usage. Unlike normal file distribution systems, HDFS uses parquet files that are well optimised for fast query performance and low I/O.

Another assignment I worked on was to analyse the correlation between network experience variables and NPS. We used Kruskal-Wallis H test and Chi-Squared test to determine if there were statistically significant differences between independent variables like network latency on NPS which is an ordinal variable.

To check the validity of the data used for our ‘Ingredients’, I learnt how to do sanity checks on sample data and cross validate across multiple sample groups. This is an important process before making any interpretations of the data.

These assignments were great practice to gain proficiency in data querying with SparkSQL as well as coding in python with packages such as Pandas and Matplotlib.

### Organisational/Industry Experience Gained from Assignments

**CHI Methodology**

To understand Lynx Analytics’ core product that it provides clients, I had to understand the Customer Happiness Index (CHI) Methodology the company uses to objectively give each customer a score based on past behaviour. By using data science to model such scores for each customer, we are able to provide focused actions the client can take to drive commercial outcomes. The outcomes that we try to model and predict are Churn probability, Average Revenue Per User (ARPU) and Net-Promoter Score (NPS).

Through being introduce to the CHI Methodology, I learnt about how consulting and analytics firms’ package and deliver their product to clients. This methodology can be applied widely to all sectors of business and would be useful in helping me frame business issues and form analytical solutions for future projects.

**Work Ethic and Communication**

Through this internship I have learnt the importance of communication, especially on the progress made or challenges faced while doing any tasks. With the lack of prior internship experience, I had not experienced the pace of working in the real world and the gravity of being accountable to more than just oneself. Certain tasks given were vague to me initially but due to a certain sense of pride I initially did not ask many questions and tried to do the task to the best of my understanding. I have learnt over the course of the internship that timely questions can make a world of difference for one’s efficiency and ability to produce effective work. It would also help you build understanding with supervisors and colleagues for the betterment of future tasks.

## Conclusions

### Summary of Work Completed and Training Received

1. LynxKite Tutorials
   1. Basic Graph Theory
2. VPN Setup
3. PySpark
4. SparkSQL
5. HDFS
6. Data Science
   1. Data distribution visualizations
   2. Correlation analysis
      1. Kruskal-Wallis H test
      2. Chi Squared test
7. CHI Methodology

### Personal Experience

As part of a data science team, we are oftentimes working in parallel, and it requires great amounts of communication to know what each party is working on. The STC project team does a daily stand-up call, led by the project manager to help keep everyone on task. I have learnt a huge amount about communicating and asking when I do not understand anything because not completing tasks within the specified deadlines could result in delays for other teammates.

An area to work on would be the ability to self-learn and rely less on my project supervisor for guidance as the project progresses and I become more familiar with the quality of work expected. This internship gave me ample opportunity to debug long lines of complex queries and develop a keen sense of troubleshooting bugs in my own codes.

Overall, it was an eye-opening experience getting to be a data science intern at an analytics company that works with large companies internationally. I had learnt a great deal about the inner workings of a data scientist as well as how to work in a team with business consultants and data engineers which is an experience that cannot be found in a school setting.