**MSc Project - Reflective Essay**

|  |  |
| --- | --- |
| **Project Title:** | Blockchain Anomaly Detection |
| **Student Name:** | **Aashish Rajendra Waghela** |
| **Student Number:** | **210557820** |
| **Supervisor Name:** | Lin Wang |
| **Programme of Study:** | Big Data Science |

In the past couple of years, the cryptocurrencies gain a huge momentum with everyone trying to explore cryptocurrencies and get a hold of it, as it served as a way to make money in a shorter duration of time with no hassles involved in buying or selling it. This momentum helped the main technology behind it to get the recognition that was missing since it started a decade back. In this case “Blockchain” being the backbone technology used by all forms of cryptocurrencies.

In general, for people to use the blockchain technology as a means for cryptocurrencies, the advantages of blockchain served as the core functionality which helped in having a secured and trustworthy way of buying, selling, and storing cryptocurrencies. Me personally being a fan of the way blockchain works and helped in having millions of transactions occurring every minute without the worry of security about the transactions. Since the time I have started to learn data science, I wanted to explore the way blockchain worked and the science used behind it. But with great technology comes risks, the two main problems faced by everyone is the problem of security and information theft. Apart from these two, comes a great risk of cyber-attacks which is also looms. But from the perspective of blockchain, it is popular because of it’s strong features like security and immutability. And being a public ledger also facilitates transparency among all the transactions occurring within the system and records being stored permanently. Blockchain is regarded as one of the most disruptive computing paradigms after the Internet. Several unique characteristics have made it a breakthrough technology for registering, verifying, and managing transactions. However, there still lies some problems within the system of Blockchain.

The main problem being the types of frauds that can happen within the system and second being the vulnerability blockchain has. To start with, malicious activities can still occur within the system. Fraudulent or malicious transactions being the most repeatedly used form of fraud. The aim behind this type of transactions is to gain undue advantage wherein it causes loss of money or cryptocurrency to the targeted victims. Thereby, malicious transactions being a point of concern among every financial industry. To solve this problem, various state of the art systems has been developed which are designed specially for the purpose of detecting malicious transactions. But there is no system which can give a set defined solution for this problem. One such solution proposed by **“M. Signorini et al. wherein he exploited blockchain metadata for malicious activity. The idea was to collect local attacks that were injected in the form of malicious transactions and later to reuse them to prevent similar types of attacks on the untainted nodes.”**

Not only in financial sectors, but the problem of anomalies being present is also prevalent in every other industry that use technology. There the problem would be, to get into the respective system and use, modify or change the metrics of the system which might benefit specific individual or an organization. To prevent this kind of attacks, anomaly detection method is used in place to stop this kind of attacks at early stage. Well, the idea to study the specific sector of blockchain for anomaly came through the coursework which I had in the second semester wherein we analysed Ethereum transaction using big data technology like spark and Hadoop. During this, we analysed number of transactions occurring every month, maximum transactions by person every in a month, top 10 miners among all the users and my favourite part wherein popular scams that had happened were found as they were already labelled as malicious transaction.

This idea kicked in to explore and find more about the anomalies that are present within blockchain. To start with that, I had to understand about what blockchain is and how it works i.e., the working principle behind blockchain. Then the features used by blockchain which makes it extremely strong against any types of attacks, what are the types of attacks to which blockchain still is vulnerable. After that what anomaly is, what can be considered as an anomaly. Different types of methods that can be used for detection of anomaly. What are the techniques used in industries and a thorough research about how anomaly detection is correlated with blockchain.

Blockchain is an interesting technology which is immune to attacks but as famously said that every system that uses internet is a hackable system. The research about the project gave new ways about how to approach a problem and how to tackle problems. Also got to learn the way anomalies are changing with the technology evolving at an extremely fast pace and new techniques used to solve the same. After getting the knowledge about blockchain, anomaly and its detection methods. The next step was to plan and execute machine learning strategies that will help in achieving better results at the same time improving my overall technical skill as well as my knowledge.

I researched about the articles on websites like medium.com, towardsdatascience.com which helped me in getting to know about the new ways which are used as well new libraries and methods used to tackle the problem of anomaly detection. One such interesting library that I came across was PyOD. PyOD is python toolkit which is designed specifically to assess the problem of anomaly detection. I read few articles which helped me to finalize my thoughts about which library to use and how to go about the problem. The next step was to identify and get a dataset which is compatible with the library I planned to use. After going through multiple datasets, I got a dataset which had Bitcoin’s transaction which I have used to get the results. The problem with most of the other dataset was that they contained time series data which I had already used for my coursework. Next step was to decide on the algorithms to be used followed by the method. K-means being used by mostly every research for anomaly detection is considered a state-of-the-art algorithm. This algorithm is used for comparing and analysing the results. Apart from this, three other algorithms have been used namely Histogram based outlier detector (HBOS), clustering-based local outlier detector (CBLOF) and Isolation Forest. Now the most important task was, how to approach the problem as multiple characteristics needed to be analysed and with a greater score.

For all the algorithms, normal implementation was not okay as it would not produce a good score, also the problem with underfitting and overfitting loomed. To solve this, iterative approach has been used where due to having multiple iterations solved the problem of overfitting. To start with the project, exploratory data analysis proved to give an upper hand in analysing most of the data. But the problem lied with the data having 30.25 million entries which made it difficult to generate scatter graphs as it will exhaust all the computing resources available on my system. To get a better hand over it, boxplots and correlation matrix gave the support in finding the types of relation between the data points (positives and negatives both). Also, the classes being heavily skewed with only 108 malicious data points made it difficult to move forward with the analysis. To overcome this problem after multiple attempts and research, I got to know about SMOTE. Wherein, it helps the algorithm by generating synthetic samples for minority data points. Without this method known, the job would be extremely tedious as finding minority data out of millions of data is no joke. Also, it would be extremely time consuming and frustrating at the same time.

The algorithms designed helped in achieving considerable amount of accuracy. But the main factor that played an important role was the time consumption used by the algorithms, also, the way accuracy was in relation with time consumption. Also, not having parallel GPU resources, the memory got exhausted while rerunning an algorithm multiple times as a set amount of memory always get used and that is not reusable. If you want to re-run it, you are required to restart the kernel, because of which pre-processing of data takes considerate amount of time. The next thing was to get a hold on the accuracy for which the hyperparameters needed to be adjusted which gave a major boost in achieving higher accuracy. To get a sweet spot where the number of iterations isn’t too large, nor the hyper parameters are absurd. The iterations after multiple attempts were set to 15. K-means being considered as the state-of-the-art algorithm for finding anomalies proved to be accurate with results due to the reason of using iterative approach and training multiple models. Which also solved the problem of overfitting. To get better hold on results the rest of the algorithms (HBOS, CBLOF and Isolation Forest) evaluation metrics have been used which described metrics like accuracy, balanced accuracy, precision, recall, F1, time. This table helped in getting the best results among all the results. One interesting result that I got from the graphs generated from the metrics was that the training time and testing time had considerate amount of difference for every algorithm. Apart from these the confusion matrix generated with correctly identified labels where the results were tested for both activities considered as malicious as well as non-malicious activity. Due to SMOTE the synthetic data points were identified correctly and the count for the same was displayed. The result achieved was up to the standard, but I think that time was an important factor that played a critical role in analysing the results. Because of the system I was using, there was a limited access to GPU due to which the estimated time of completing more algorithms was not achieved. The main aim was to implement more algorithms and then to do a comparative study which would enable more about the workings of algorithm. But due to restriction and time constraint, only 4 major algorithms were implemented.

The next steps in the future would be to build a system using these algorithms which will enable us to identify any malicious transaction happening at the user node. Finally, to conclude the research, the project enabled to understand about new terminologies as well new technologies which proved to be helpful for me in the coming years. As the project helped me to get an in-depth knowledge about the techniques as well as the methods that are used among some of the most renowned companies. I had never thought that I would get to know something as interesting as blockchain and anomaly detection by completing my research project. Also, this boosted my overall confidence in approaching a problem and how to research through a problem and find useful insights from them.

References:

1. Jennifer J. Xu, “Are blockchains immune to all malicious attacks?” in Financial Innovation (2016) 2:25.
2. Sirine Sayadi, Sonia BEN RJEB and Zied CHOUKAR “Anomaly Detection Model Over Blockchain Electronic Transactions” in 15th International Wireless Communications and Mobile Computing Conference (IWCMC), Jun 2019.