Heinz 95-845: Classification and Time Series Analysis of Bitcoin Users

Simran Handa SHANDA/shanda@andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Prakhar Mishra PRAKHARM/prakharm@andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Malvika Singh MALVIKAS/malvikas@andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

1. Proposed Analysis and Likely Outcomes

In our project, we are trying to look at bitcoin pricing, as well as the intricacies of blockchain that come along with it. By using FigShare data[1] that details transaction data for customers, our project aims to identify average timings of transactions, after which we would also like to look at user IDs in order to know whether the users are serious traders or hobbyists engaging in trading. We would also like to know if the cryptocurrency asset is a bubble or not by analyzing if the consumer base for cryptocurrencies is actually increasing or if the same users that are increasing the frequency of their transactions.

2. Importance and Potential of Analysis

Ever since bitcoins (and cryptocurrency in general), first made their foray as a form of money transfer, there has been a lot of confusion regarding how they can be adapted and integrated with the existing global financial markets. With our analysis we hope that we will be able to understand crypto transactions a little better, which would serve as a small step towards a much better goal that could revolutionize the payments industry.

3. Evaluation Measures

Our problem is primarily a classification problem, so some of the possible metrics that we would like to use are –

- Accuracy / Precision / Recall
- F1 Score
- Log Loss / Binary Cross-entropy

- Categorical Cross-entropy
- Area Under the Curve

We would specifically like to look at the number of False Positives and False Negatives, considering that both have potential downsides associated with them. In this case we believe that false positives would be more costly as it would lead to a company expanding resources in a market which is not financially profitable.

4. Relevant Work

During our literature survey, we gathered information about three main works done in this field. These are as follows:

- Research titled *'A Bayesian approach to identify Bitcoin users'* by Juhász, P. L., Stéger, J., Kondor, D., and Vattay, G. (2018) uses supervised learning methods to identify bitcoin users and predict their potential to participate in more transactions.
- Work by de Souza, M. J. S., Almudhaf, F. W., Henrique, B. M., Negredo, A. B. S., Ramos, D. G. F., Sobreiro, V. A., and Kimura, H. (2019) titled *'Can artificial intelligence enhance the Bitcoin bonanza'* uses Support Vector Machine and Artificial Neural Networks to predict cryptocurrency prices.
- Paper by Moore, Tyler, 2017 on *'Replication Data for: Revisiting the Risks of Bitcoin Currency Exchange Closure'* uses logistic regression to identify the major causes of bitcoin frauds.

5. Our Contribution

Our work is novel in the sense that it studies the behavior patterns of current users in a time series analysis. We focus inward to capture the trend in consumer behavior over time. The objective is to identify and classify the seasoned(risk neutral) versus the hobbyists(risk averse) people from the data and observe the variation in their transaction footprints over time. This will benefit financial institutions trying to expand their market to bitcoin and to understand when there is a spike in consumer behavior and look at whether the consumer base is expanding or not.

6. Data Description

The dataset comprises of 5 sub datasets. These 5 include:

- **Capture dataset (1.5 GB)** : This sub dataset comprises of parameters like sender_ip indicating masked IP address from which the message was sent by the user, timestamp for unix format time stamp, transaction_id (bitcoin transaction ID), Monitoring_client_ip reflecting masked IP address of the monitoring client which recorded the message

- **Address Lookup dataset (60 MB)**: This sub dataset[2] comprises of following parameters: Sender_address - masked input (sender) Bitcoin addresses of the transaction, transaction_id (bitcoin transaction identifier)

- **User Grouping(20 MB)**: This sub dataset[3] comprises of following parameters: Address_id - bitcoin address , timestamp - user identifier that owns the bitcoin address

- **Accepted Pairings(261 KB)**: This sub dataset[4] comprises of the following parameters: user_id - user identifier,ip_address which is masked IP address that is owned by the user, probability indicating probability of pairing.

- **Identified Transaction(821 KB)**: This sub dataset[5] comprises of the following parameters: sender_userid - originator of the transaction, receiver_userid - target of the transaction, timestamp - unix timestamp of transaction

For the classification analysis, we plan to create a variable called isHeavyUser which will take the binary value 1 if the number of transactions performed by that userID is greater than a threshold value. We plan to develop this threshold value by considering the median of the number of transactions for all userIDs. This created variable isHeavyUser will be the outcome Y. The treatment U will be the transaction_id (indicating that the user has performed a bitcoin transaction). The population W will consist of user_ID field and all the observations in that field. The covariates will be probability which indicates the probability of pairing and timestamp/month (month will be an extracted variable from the unix timestamp). The time series analysis using neural networks will have the outcome as the probability of transaction in the next time period.

7. Study Design, Data Pre Processing and Proposed ML Techniques

We are using panel data involving observations for multiple variables over time. The study design we will be using is longitudinal. This is because we are using observations of same parameters over different periods of time.
The preprocessing involves joining the datasets based on primary key transaction_id. We plan to merge capture and address dataset giving information about the when the transaction was made by a particular user id. So this will help us study the frequency pattern of the users. The other two datasets will be merged using userID and these two datasets combined contains information about probability of future transaction. The dataset available to us is in txt format which also needs to be converted to csv in this step. The time series analysis using neural networks will have the outcome as the probability of transaction in the next time period. The preprocessing also involves creation of output variable isHeavyUser based on proposed threshold level as discussed earlier. Machine Learning methods that we plan to use involves supervised classification like logistic regression and time series analysis will be done using neural networks. As described in Section 1.1, we believe that this project is novel as it aims to study patterns of investing by looking at the transactions on bitcoin. This study will allow portfolio creators to know how newer types of stocks, such as cryptocurrencies, can be incorporated into customers' portfolios. We believe that we can use these datasets and this form of study for this project as our dataset is large and varied

enough, showing us several different users and their transaction patterns, to allow us to gain useful insights about different levels of bitcoin investment.

8. Possible Limitations of this Study

- **Lack of detailed personal features**: Here, we see few columns detailing demographics about the user, such as age, ethnicity, or gender. This makes it more difficult for us to gather data based on extremely specific factors to form an investment portfolio for different types of users and we therefore must rely on factors such as time and frequency to form the bulk of the data used for our portfolio creation.
- **External factors**: While we have data on the number of times and the rate at which users use and conduct transactions through bitcoin, which allows us to approximate the state of the world economy, we have little concrete data on other outside factors that may influence their buying/spending patterns, such as economic downturns in their home country only, personal crises, or competing cryptocurrencies.

9. Users of Analytic Pipeline

We would like to develop our pipeline in such a way that it is put into use by banks in order to develop portfolios for consumers, dependent on certain information given by their customer, such as their willingness to take risks and their interest in investing in newer versus more traditional stocks. More traditional banks such as Goldman Sachs and J.P. Morgan are beginning to invest in bitcoin and other cryptocurrencies and this analysis will allow them to help their customers.

References

[1] Steger, Jozsef (2018): Capture dataset. figshare. Dataset.

[2] Steger, Jozsef (2018): Address lookups. figshare. Dataset.

[3] Steger, Jozsef (2018): User grouping. figshare. Dataset.

[4] Steger, Jozsef (2018): Accepted pairings. figshare. Dataset.

[5] Steger, Jozsef (2018): Identified BC transactions. figshare. Dataset.

[6] Steger, Jozsef (2018): Bitcoin transactions. figshare. Media.