ML Project Proposal
Title: An AI Perspective on the Cryptocurrency Market: Predicting Pump-and-Dump Schemes
Team Members: Cyrena Burke and Daniel Olaes

1 Introduction

At first, the idea of hashing and storing completely digital currency in a digital ledger was nothing more than a gimmick to the public. However, over the years, as more and more cryptocurrencies were developed and released, the idea of digital currency gained validity in the public eye and began to grow. Nevertheless, there have been several instances of scammers manipulating cryptocurrencies by inflating the price through high-volume transactions, then "pulling the rug out," allowing them to accumulate large amounts of money at the expense of shareholders who fell for the false hype. Therefore, by utilizing a dataset of the hourly historical OHLC + Volume data for several cryptocurrencies, we decided to implement a learning system to identify and predict crypto market pump-and-dump schemes.

2 Problem

Given the open price, high price, low price, close price, and volume amount of one of our cryptocurrencies, our system should be able to utilize historical data to identify if that currency is experiencing a dramatic fluctuation in volume and price at that given time. By comparing that instance to previous rises and falls in the crypto market, the system should be able to predict whether the currency is experiencing organic price shifts or intentional market manipulation

3 Input

The input will be an instance of one of the 6 cryptocurrencies during a given hour, including details like the open price, high price, low price, close price, and volume amount.

4 Output

The output will be a classification of the behavior exhibited in the instance: organic value growth or fall, market manipulation pump or dump, or a complete market crash (5 total classifiers).

5 ML Technique

We decided to utilize Naive Bayes classification because predicting market changes is a multi-classification problem. In addition, we aim to train our model with a semi-supervised approach by identifying the different classifications in a small portion of the data while a majority remains unlabeled. To develop our model, we will use an 80/20 training/test split, where the training set is 80% of the data proportionally taken from each of the six cryptocurrencies and 20% of the data is the testing set, allowing us to test each cryptocurrency equally.

6 Innovation

In researching the topic of cryptocurrency, we found multiple papers about our original topic of using machine learning to predict market activity, primarily crypto booms and crashes. However, we did not see a lot of discussion about intention market manipulations, such as pump-and-dump schemes. Therefore, we chose our 6 cryptocurrencies very intentionally: 2 accurately demonstrate generally stable market behaviour, BTC (Bitcoin) and ETH (Ethereum), and 4 that were the victims of pump-and-dump schemes during the 2017-2018 time frame available in our data, DOGE (Dogecoin), XRP (Ripple), XVG (Verge), and NXT (Nxt).

7 Dataset Details

This dataset details the hourly historical OHLC + Volume data for BTC (Bitcoin), ETH (Ethereum), DOGE (Dogecoin), XRP (Ripple), XVG (Verge), and NXT (Nxt).

Instances/Rows: 31621 instances

Fields/Columns: 9 fields (Unix Timestamp, Data, Market Symbol, Open Price, High Price, Low Price, Close Price, Share Volume, USD Volume)

The downloadable dataset can be found under the following URL:

https://www.cryptodatadownload.com/data/hitbtc