Harvard John A. Paulson School of Engineering and Applied Sciences

IACS Institute for Applied Computational Science

Project Outline Document

Team	David Assaraf, davidassaraf@g.harvard.edu Connor Capitolo, connorcapitolo@g.harvard.edu Tale Lokvenec, talelokvenec@g.harvard.edu

Problem Definition and Proposed Solution

Forecasting Crypto Exchange Prices - The current state of the crypto market is extremely volatile. Due to lack of experience/exposure from traditional actors (the crypto market is still in its early adoption stage), there is a lack of systematic investment strategies in the crypto environment; therefore, there is an opportunity to extract value from an accurate prediction of the price dynamics of exchanges. Focusing on different platforms (Binance, FTX, KuCoin), we will aggregate various types of data in order to build a predictive model of the price dynamic of exchanges. The objective is to produce real-time predictions on data provided by the different platform APIs. Our aim is to tackle three different objectives:

- bridging the lack of structure dealing with crypto exchanges in building a scalable and modular database architecture that will gather various features from different platforms for exchanges; we plan to initially start with the Binance API and build from there
- building a predictive ML/DL model that will enable us to gain insights as to how the market is evolving over time in order to inform trading decision making

The final product will be a web application that has a drop-down menu which allows users to choose which exchange they want to examine. Upon clicking the particular exchange, an additional page will appear that shows exchange rate predictions in bold for the next minute, hour, and 24-hour time period, along with a line graph that shows minute-by-minute predictions over the next 24 hours that individuals can scroll over. These predictions will be updated every minute to reflect the additional data received.

If time permits at the end of the project, we would like to add additional data that may be useful in predicting exchange rates, such as news articles or financial market data.

Dataset(s)

The dataset being used is composed of the different features we can extract from the different APIs' specific exchanges. For now, we will focus on historical data involving the exchange rates in order to train our large scale Deep Learning models. Regarding the Binance API specifically, we have 1,612 exchanges with the first data point being collected on August 15, 2017. The current Binance features with available history are called “candlesticks”. Candlesticks are aggregated trades over a certain period of time with features summarizing the trades that happened for the specific period. Here is an overview of the candlestick features:

Candle Features	**Features Description**
Open Time	Candle Open Time
Open	Open Price in Quote (Secondary) Asset Units
High	High Price in Quote (Secondary) Asset Units
Low	Low Price in Quote (Secondary) Asset Units
Close	Close Price in Quote (Secondary) Asset Units
Volume	Total Trade Volume in Primary (Base) Asset Units
Close Time	Candle Close Time
Quote Asset Volume	Total Trade Volume in Quote (Secondary) Asset Units
Number of Trades	Total Number of Trades
Taker Buy Base Asset Volume	Taker (Matching Existing Order) Buy Base Asset Volume
Taker Buy Quote Asset Volume	Taker (Matching Existing Order) Buy Quote Asset Volume
Ignore	Safe to Ignore

Using an update frequency of 1 minute, the size of the dataset for one exchange [BTCUSDT] is 0.5Gb. Working with ~1000 exchanges, we expect to have a total database of ~500Gb.

Models Being Considered

We are planning to initially approach this as a time series problem. We will use simple time series models, such as Autoregressive Integrated Moving Average Model (ARIMA), as a baseline in order to predict the price of an exchange given its history. We will then pursue more advanced time series models, such as Long Short-Time Memory (LSTM), and compare its performance to the baseline. One issue we might encounter is the update frequency of the model. Since we want to make real-time predictions, we will need to retrain/fine-tune the model when being exposed to the latest market information. This might be compute-intensive, so a clever approach will be valuable here. One idea we are currently discussing is to perform feature engineering to transform the time series data into tabular data and use Random Forest or Boosting algorithms. Depending on the results from the time series and tabular data modeling as well as the project timeline, we may extend the scope of our project to develop a trading agent. The idea is to combine the best-performing time series data/tabular data model with a Deep Reinforcement Learning (DRL) model to produce an optimal trading agent.

Project Timeline

Week Ending	Tentative Milestone or Goal
2021-09-24	● Project set up ● Explore Binance API ● Get used to python-binance library and ways to query the API ● Explore the features available on the binance API that will get signal
2021-10-01	● Create local python script in order to query historical data from Binance API
2021-10-08	● Deploy python script on the cloud (AWS) ● Set up database infrastructure on the cloud with historical data ● Work on local Websockets in order to constantly update our cloud database
2021-10-15	● Deploy the Binance websocket to the cloud ● Make sure that the database is able to scale up (anticipate ~500GB of data across 1600 exchanges)
2021-10-29	● Work on data pipelining: feature creation ● Devise modelling strategy: train/test/split ● Identify exchanges we want to model
2021-11-05	● Refine model selection ● Start model training ● Deal with cost issues in terms of model training
2021-11-12	● Work on LSTM model, model training, model validation ● Start working on real-time prediction: devise re-training strategy, update frequency
2021-11-19	● Finish real-time prediction pipeline
2021-11-26	● Start working on the App
2021-12-3	● Finish up work on web application
2021-12-10	● Create Blog post and video