About Us.

The team is comprised of 4 crypto enthusiasts (1 long term believer and 3 recent converts), who appreciate the simplicity (while concurrently marvelling in the complex mathematical sophistication underlying the cryptography) and believe in the transformative possibility of the technology. We started [mydiamondhands.io](http://mydiamondhands.io) with the mission statement of: Educating retail investors on the transformative opportunity of crypto and providing them access to a arbitrage model which serves to build confidence and engagement in crypto currency markets.

To bring our mission to fruition, we have identified 3 guiding principles 1) Education over Enrichment 2) Simplicity over Sophistication 3) Community over Competition. These principles are used to guide decision making, to ensure that we retain integrity, consistency and never lose focus on our goal of positively contributing to the crypto community.

Our Approach

To bring our mission to life, while adhering to our guiding principles we needed to develop a plan on how to execute. As a team we brainstormed focusing on our passions, and arrived at a series of goals we believed possible, that would result in a positive contribution to the crypto community. Those team goals included;  1) Creation of a one of a kind data set (to be published on git hub), which consolidated features ground in theoretical principles which could be used to excite, educate and ultimately populate our ML model 2) Creation of atheist 1 unique ML Model, which could be used to profitability implement a low risk investment trading strategy. 3) Create a communal dialogue, share the work, findings, insights, research and development work undertaken with the sole purpose to educate, empower and promote continued positive and intellectual dialogue in the crypto community.

While creation of a data set was established as our first deliverable, we felt that understanding the model/ trading strategy would inform the requirements and structure of the dataset, as such started with conceptualization of a strategy. While the goal was to develop and implement a profitable strategy, it became apparent that expectations could be a challenge, specifically the gains of holding crypto over the past 20 years have been exponential for pioneers who bought and held, with stories of overnight millions/billions common place. However, during the past 5 years volatility in pricing has introduced considerable risk, and while many have implemented successful trading strategies, others have lost considerable fortunes in search of riches. As such it is critically important to highlight the inherent risk with crypto as a both a short term and long term asset (given run up in valuations). Given this risk profile, the group sought a primary strategy which reduced risk, which could be attractive to help entice, educate and inform traders, such that before they exposed themselves to much riskier strategies, they possessed experience, knowledge and undestanding. Specifically, as our approach is meant to if possible eliminate risk, we emphasize that our comparable benchmarks will be risk-free or near risk-free assets such as TD, GICs and Bonds.

The team explored strategies which met this criteria (Risk Free, or Near Risk Free)  and identified two primary trading strategies, which provided investors the opportunity to eliminate holding risk (should they decide that holding a particular crypto currency is outside of their risk profile) and would meet our criteria. The first strategy leverages the concept of high frequency trading, while both exciting and profitable, this strategy is distinctly outside the scope of interest, as it fails two or our guiding principles; Complexity; it requires expensive, high performance hardware, and Competitive; it is solely based on being faster and more precise that fellow traders ( in certain strategies you explicitly exploit less effective strategies). The second strategy, which is our preferred strategy utilizes the development and testing of strategies which utilized ML Models to identify instances when circumstance dictated in the near immediate future the expectation of arbitrage opportunities, to enable traders to both understand what was happening, and execute a profitable strategy.

Development of Dataset

When starting the trading strategy process our team sought to understand what might be of relevance and interest, both in the pursuit of building a prediction model, and in the development of ancillary assets to assist in the development of learning, education and promotion of insights related to crypto. As such we consulted a number of theoretical sources, which included strategies and papers related to crypto in general, the underlying technology, crypto trading, crypto investing and crypto arbitrage (LINK TO DICTIONARY). In reviewing the literature it was apparent that considerable work was being devoted to the development, testing and implementation of strategies, however the inputs were manifesting into drastically differing outputs varying in beliefs, expectations, reasonable goals, strategies, implementation and execution.

In order to develop a dataset several key components were required, the most foundational we needed to identify a crypto currency (or pairing) with which to focus, and then exchanges where trades could be compared, monitored and executed. Several factors were considered in the decisioning processes, the most foundational being, data availability, liquidity, volume, volatility, and platform equivalency.  Starting with Data Availability, our foundational dataset and model requires a consistent, high quality and readily available data source. Specifically, as we planned to capture and identify trader related behavioural components, the need for accurate, consistent and granular information is high. Next, we needed to insure that whatever markets we identified were liquid, in the sense that transactions continued to frequently occur, had real time prices which were reflective of prevailing market conditions, and when we ultimately deploy our model, we will be able to execute trades to capitalize on our strategy (specifically there are numerous platforms and exchanges were instantaneous transaction implementation is not possible). As supported in the literature, we identified that while a asset volatility is required to support the continued availability and development of a effective market, that extreme price volatility is undesirable, both in the development of a long term valued asset (and stable currency), but most importantly, extreme volatility in the context of a arbitrage model can be catastrophic, presenting risk we do not seek to introduce. Lastly, equivalency is needed to ensure that the underlying platforms where trades will be executed are identical (or near identical). Differences in platforms (whether real or perceived) can result in pricing differences, which could artificially suggest the presence of arbitrage, but simply be the manifestation of operational risk or performance issues. As an example, in the past 10 years a number of previously well respected trading and storage platforms have failed for various operational, risk, fraud related reasons including; FTX (2022), Mt. Gox (2014),Bitfinex (2016),QuadrigaCX (2019),Cryptopia (2019),Celsius (2022),BlockFi (2022), and Zipmex (2022).

Ultimately the decision was made to focus on the currency pairing on Uniswap pools, and the specific currency pairing of WETH/USDC. The Uniswap platform was identified as the ideal platform as it is a transparent, open source, decentralized protocol, which is clearly transparent, consistent and removes the vast majority of counterparty risk associated with individual exchanges or companies. Additionally as it is decentralized and open source, it is widely accepted throughout the crypto community, resulting in familiarity, trust, consistent performance. Additionally, as this is a standardized technical protocol, it is utilized across a wide range of currencies, pairs and pools, as such the foundational components of our data pipeline can be easily extended, allowing for leverage and scale in the pursuit of model creation. WETH and USDC were chosen distinctly, with WETH being chosen for several desirable properties, and best in class technical innovations related to the Ethereum Blockchain which were not directly available in a number of peers (components which are highlight throughout the documentation uniquely). USDC was chosen specifically for its stability, USDC is a stable coin (which means based by US dollar currency on a 1:1 ratio) as such we are reducing operational risk to a single currency, and given its 1:1 value with the US dollar, it makes comparison pricing very easy to interpret and comprehend, which we believe a very desirable asset for new and beginning traders (as the complexity of the underlying technical is considerable). Additionally, as USDC is within the currency pairing, there remains an opportunity for an individual to liquidate their position at any point in time, to completely eliminate exposure related to holding a crypto asset.

Having identified the currency pairing and platform, the next activity was related to sourcing the data. Given the choice of dataset, there were dozens of sites which publish real, or near real time information related to these pools, and could serve as a great source. The group focused on a particular API which was available on [TheGraph.com](http://TheGraph.com), which provided the foundational information necessary underlying transaction data via the API at no cost (Please refer to Appendix \_\_ ). While this provided the necessary foundational information, through development a number of additional variable were identified as important (or potentially important requiring additional exploration) and a supplement source was identified from [Etherscan.com](http://Etherscan.com) which provided information related to gas limits and historical trader activity. It is noted that while the initial dataset was built on this API, it subsequently stoped working, as the host has taken it down, fortunately as per our requirements there are hundreds of viable alternatives, with the current model being serviced by \_\_\_\_\_\_\_\_ (available in Appendix).

Having now selected our strategy, and obtained our foundational data source, it was time for the initial EDA and feature engineering (please refer to the data dictionary for specific datum level analysis). The group started with a simple summary, aggregation and interpretation of the dataset, which included looking at Value Counts of top 5 occurrences of all variables, counting the total number of unique instances, missing, zero values, NAs, and statistical components for numeric variables (Mean, Median, Max, Min, Sum, Std). The observations are stored in the GIT Repo for all variables (LINK!!!!! Include comment related to Interpretation if can not be included in HTML).  These observations led into the creation of a number of new elements, which were variants and transformations which we believed would perform better within the context of the model, specifically noting about some of the unique data formats which are common in their Ethereum ecosystem but not prevalent elsewhere (ie. Wei /Gwei ).

At this point we were presented with a unique challenge, specifically how to structure the model to enable comparison analysis between pools. While several approaches were considered, the datasets were merged on time, with an outer join, ensuring that all transactions were included within the dataset and the transactions from each unique pool would be opposite its closest comparable. Unfortunately as each pool is distinct there is not a natural primary key, as such many transactions are left without a comparable as hundreds of transactions can occur within a particular pool without a corresponding contra entry. In order to bridge this shortcoming and ensure that every transaction is relative to its most recent pool entry, we have forward filled the dataset, such that the previous transaction within the pool is available for comparison. While this approach resolved challenges for modelling, it created some unique feature engineering challenges, specifically through the introduction of considerable duplication. This comparison was handled with the introduction of a unique flag related to unique transaction ( Tran1Unique,Tran2Unique).

It was at this point we were able to engineer critical features required to create a base line prediction, specifically utilizing the available pricing and fee data (\_\_\_\_, \_\_\_\_\_ Gas Used, Gas Price, and Uniswap Transaction Fees) to identify prevailing differences in pricing at the time of transaction execution, and creating the target variable, identifying whether arbitrage was possible at the time of the transaction (percent change, swap\_go\_nogo).

With a primary data set in place, the group sought to explore additional features which could be added into the model. As crypto investing is still in its infancy, consensus with respect to optimal feature selection is still hotly debated. Instead the group took a broader perspective and looked towards financial markets, where investors have been purchasing and selling financial assets for centuries. Technical analysis is a long standing discipline which focuses on price, volume, frequency and directional indicators over time to predict movements in asset pricing, with thousands of generally accepted and utilized features utilized to predict to stock, currency and commodity movements. Borrowing from this theory, we have engineered a series of activity based features which attempt to capture, User, Pool, Daily and Market movements across varying periods of time. The rational for inclusion is twofold, 1) capturing specifically what people know into a model is not practical, however capturing peoples behaviour in the form of their activity is relatively simple (and given financial motivation, how people might be the best proxy for what they know), 2) the best indication of how the market is likely to react to a particular circumstance, is likely similar to how it has acted in the past.

As we have spoken crypto is nascent in its existence, however the financial markets have undergone over 200 years of booms and busts  (Panic of 1837 caused by a crash in cotton prices, Depression of 1873 caused by the collapse of the Vienna Stock Exchange, Panic of 1907 caused by the collapse of the Knickerbocker Trust Company, Black Monday which served as the start of the great depression of 1929, the Oil Crisis of 1973, the savings and loan crisis of the 1980s, Asian Financial crisis of 1997, Long Term Capital Management 1998, the  Dot Com Bubble  2000, and the Global Financial Crisis of 2008 )and these trials, theory and practice have adopted and adapted.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generatedA screenshot of a white background

Description automatically generatedA graph of different colored squares

Description automatically generated with medium confidence