The University of Hong Kong

FITE 4801 Final Year Project (2024-2025)

FYP24061: Bitcoin Pricing Insight using ML and NLP

Project Plan

Chan Ka Ho (BASc FinTech)

Ku Pak Ho (BASc FinTech)

Lee Chun Hang (BASc FinTech)

Supervised by Professor Chow Kam Pui

1. Project Background

Bitcoin have evolved from a decentralized money transfer solution to a multi-purpose financial asset traded globally. There is a growing number of investors, speculators, and institutional professional actively involved in Bitcoin trading market (Winotoatmojo et al., 2024). One of the primary objectives of these market players is to maximize returns and generate alpha, which often requires detailed research to gather valuable insights for developing effective trading strategies. By leveraging the information advantage, they can increase their chances of executing successful trades.

As a leading cryptocurrency with the highest market capitalization, Bitcoin often shows a different price movement behaviour compared to other asset class in the financial market, and it is widely considered as a speculative investment (Karau, 2023). Nevertheless, it shows certain degree of relationship with S&P500 index, oil price, and economic variables (Chen, 2023). To discover features that can predict Bitcoin's price movements, data used for training and signal generation does not limit to market data, but also involving alternative data such as news articles and social media posts (Frattini et al., 2022). With more data and advanced technology available, researchers have been accessing various quantitative trading strategies, ranging from regression models (Chen et al., 2020; Chen, 2023) to more complex deep learning algorithms (Nair et al., 2023; Parente et al., 2024) and natural language processing techniques. (Li et al., 2020; Parente et al., 2022; Kaur & Sharma, 2023; Girsang & Stanley, 2024)

Building on existing strategies discussed in literatures, this project aims to integrate the most effective trading strategy from multiple sources and re-evaluate models in light of the evolving economic condition, regulations and investors’ behaviour. The report begins with a review of relevant studies, followed by our project objectives and methodology. Towards the end of the document, a tentative project schedule is proposed.

2. Literature Review

Quantitative trading strategies for Bitcoin pricing have been explored in many areas. Starting with regression models, Chen et al. (2020) suggested using a dataset with high-dimensional features and applying logistic regression is an effective method to predict price for low frequency trading. Chen (2023) further analysed a dataset with 47 variables and utilized random forest regression to determine the importance of each factor in predicting Bitcoin prices. The performance of these method showed a high level of accuracy before 2018 but failed to accurately predict Bitcoin price surge in 2018 and 2021.

Some complex models using neural networks are effective in predicting Bitcoin price. Nair et al. (2023) focused on using Bitcoin historical trading data to compare several deep learning algorithms for time series forecasting, including recurrent neural network (RNN), long short-term memory (LSTM), gated recurrent unit (GRU), bidirectional long short-term memory (Bi-LSTM), and 1D convolutional neural network (CONV1D). Using evaluation metrics, including root mean squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), and r-squared score, they found out the best model is LSTM, followed by Bi-LSTM and GRU. Besides, Parente et al. (2024) developed a fine-tuned multi-layer perceptron (MLP) to classifying market signals into Buy, Hold, and Sell categories. This AI model added explainability by weighting features contributing to the model's predictions using SHAP (Shapley Additive exPlanations).

For text-based data, like social media posts and news articles, a different model is required for machines to process human languages. Raheman et al. (2022) evaluated 21 different text-based models for Twitter and Reddit sentiment analysis. They concluded that Aigents model and FinBERT have the highest accuracy. Main difficulties they faced were sarcasm and idioms, which are often misclassified by models.

Since much of the literature focusing on specific areas, we think it is important to integrating models from different in perspectives, so that we can accurately predict price movements in a comprehensive manner. Also, comparing results across different studies can be challenging because different data, methodologies and evaluation criteria are used.

3. Project Objective

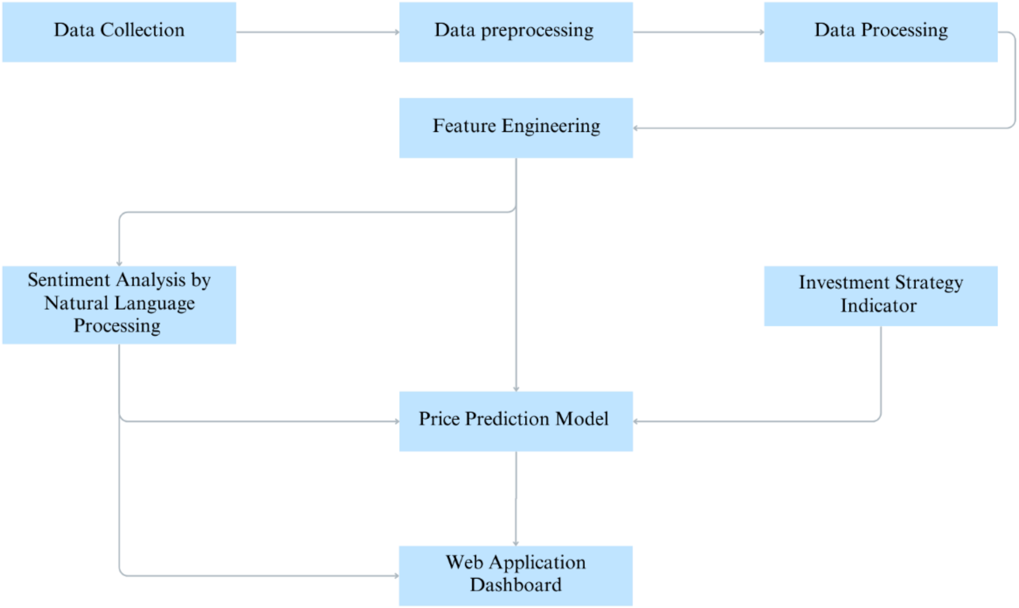
Our investment goal is to achieve capital accumulation in the Bitcoin market through a multi-strategy approach. Leveraging data, technology, and research, we take smart risk with the information advantage and aim to deliver sustainable return regardless of market condition. We further break down our objectives as follows.

- Objective 1: We use pricing model to capitalizes on short-term price fluctuation and long-term trend. Using machine learning, the model learns patterns from the past data and generate profits by buying low and selling high.
- Objective 2: The price prediction model also incorporates views and opinions discussed among investors communities on social media and news outlets. We will first access the significance of investor sentiment. This alternative data can then be used to assist the model to predict price trend based on the results of sentiment analysis.
- Objective 3: The model exploit price movements in response events, such as new Bitcoin project releases, regulatory announcements, technological updates, macroeconomic events. The pre-event strategy involves researching market rumours, assessing the likelihood of these rumours materializing, analyse sentiment. We act before its formal announcements and harvest profit if the rumour is true. At the time of the event announcement, price reaction model is used to adjust price prediction. For post-event strategy, the model continues to identify opportunities for trend continuation and mean reversion, as the market may initially overreact or underreact to the event.

For each model, the performance will be evaluated using common evaluation metrics used by researchers and the industry. Using the results, strengths and weakness of each method will be discussed, which will be used to determine the most effective model for our case. Lastly, we will create a dashboard to display the results of the chosen pricing model. This serves as a user interface for traders to get Bitcoin pricing insights.

4. Project Methodology

This section illustrates the methodology for various stages of the project. The flow of the project is demonstrated on the diagram below, outlining subsections on data collection, data processing, NLP analysis, price prediction models, metrics evaluation and web application dashboard.


Data Collection
Data preprocessing
Data Processing
Feature Engineering
Sentiment Analysis by Natural Language Processing
Investment Strategy Indicator
Price Prediction Model
Web Application Dashboard

4.1 Data Collection

The project aims to predict the Bitcoin price from the market condition and the sentiment of investors. The cornerstone of our project is the vast amount of reliable information regarding the Bitcoin. In order to collect the enormous amount of data from various sources on the Internet for analysis, different approaches are adopted based on the type of data. The type of data includes:

1. Market Data

   The market data consists of Bitcoin-related data and data of other assets.

   The market data of Bitcoin includes the open price, the close price and the volume of Bitcoin in various frequencies and the data of other assets includes the open price, the close price, the volume, S&P index, gold spot price, oil price and ETH price. The data of other assets is used to analyze the current market condition for analysis on relationship between market condition and Bitcoin price.

2. Social Media Post Data
   The social media data includes the title, the content and the traffic of posts regarding Bitcoin from Weibo, Reddit and X (previously Twitter) in different frequency. Bitcoin's subreddit and follower on X has around 7 million Bitcoin people, this shows that both platforms have a significant influence in the community. These data are essential for us to analyze the sentiment of both Chinese and English Bitcoin community in global.
3. News Data
   The news data includes the title and content of posts regarding Bitcoin from crypto platforms, including Coindesk and Cointelegraph, and traditional news platform, including Google News and Bloomberg. The news data provides insight on the regulations and market condition of Bitcoin.

To collect an extensive and diverse dataset, two major approaches will be adopted in our project:

1. Application Programming Interface (API) Approach
   - Market Data

Bitcoin-related data will be collected using the Binance API. The Binance API allows us to collect minute level market data in a fast and accurate way. For data of other assets, we will be using the Yahoo! Finance API to collect accurate and comprehensive market data of different assets and indexes.

- Social Media Post Data
  We will be using Weibo API to collect social media post data of Chinese Bitcoin community and Reddit API and X API to collect social media post data of English community.
- News Data
  To gather the news data from crypto platforms, we will be using Crypto News API. The Crypto News API allows us to collect news on cryptocurrency news platforms like Coindesk, Cointelegraph and Crypto.new.

2. Web Scraping
   - News Data
     In order to collect news data from traditional news sources, we will be using Beautiful Soup to scrap data from Google News, Bloomberg and Forbes. By gathering news data from traditional news sources, we can have a comprehensive picture on the market condition of Bitcoin.

4.2 Data Processing

After data collection, the tremendous amount of data, it is very important to perform data cleaning and processing before passing into the models to ensure accuracy and efficiency of model. At this stage, we will first remove the corrupted data, missing values and inaccurate information from the dataset, we will also analyze the outlying data and evaluate the accuracy of the outlying data to ensure the accuracy of the dataset. After data cleansing, we will perform tokenization and lemmatization for textual data using Spacy and Natural Language Toolkit.

4.3 Natural Language Processing (NLP) Analysis

To analyze the market condition based on news and investor sentiment based on social media posts, we will make use of FinBERT model to analyze the sentiment information of textual data. The FinBERT model outputs the sentiment of news and social media posts. The FinBERT model helps us to convert textual data to sentiment and evaluate the relationship between market sentiment and the price of Bitcoin.

4.4 Price Prediction Models

Commonly used momentum trading technical indicators will be passed into the model, namely simple moving average (SMA), moving average convergence divergence (MACD), relative strength index (RSI), and momentum indicator (MOM).

After collecting the sentiment of market, the textual data and market data will be passed in various models together with technical indicators to perform price prediction and evaluate the performance across different models and frequency of data. The following models will be used:

1. Statistical models such as Auto-Regressive Moving-Average Model (ARMA)
2. Machine learning models such as Logistic Regression (LR), Random Forest Regression (RFR) and Support Vector Machine (SVM)
3. Deep learning models such as Recurrent Neural Network (RNN) and Long Short-Term Memory Network (LSTM)

4.5 Metrics Evaluation

In order to compare performance of various models, we will use Root-Mean Squared Error (RMSE), R Squared value and Mean Absolute Scaled Error (MASE) to evaluate the performance of models on price prediction. RMSE shows the overall average prediction error.

R^2 illustrates how fit the model is and MASE compares the performance of model and naïve forecast (Plevris et al., 2022).

4.6 Web Application Dashboard

Lastly, the results of models will be displayed in web interface. The frontend environment will be developed based on React.js and the backend environment will be developed based on Next.js.

5. Project Schedule and Milestones

Time	Schedule
Sep 2024	Literature Review
Sep 2024	Project Plan Documentation Project Website Configuration
1 Oct 2024	**Deliverables of Phase 1:** • **Detailed Project Plan** • **Project Website**
Oct - Nov 2024	Data Crawling and Data Cleaning
Nov 2024	Models Implementation
Nov - Dec 2024	First Stage of Model Evaluation and Fine-tuning
Dec 2024	Web Application Prototype Development
Dec 2024 – Jan 2025	Interim Report Documentation Presentation Preparation
13-17 Jan 2025	**First Presentation**
26 Jan 2025	**Deliverables of Phase 2:** • **Preliminary implementation** • **Detailed interim report**
Feb 2025	Second Stage of Model Evaluation and Fine-tuning
Mar 2025	Final Implementation and Fine-tuning Dashboard Development
Mar – Apr 2025	Final Report Documentation Presentation Preparation
21 Apr 2025	**Deliverables of Phase 3:** • **Finalized tested implementation** • **Final report**
22-26 Apr 2025	**Final Presentation**
30 Apr 2025	**Project Exhibition**

References

Chen, J. (2023). Analysis of Bitcoin Price Prediction Using Machine Learning. *Journal of Risk and Financial Management, 16(1), 51.* https://doi.org/10.3390/jrfm16010051

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics, 365, 112395.* https://doi.org/10.1016/j.cam.2019.112395

Frattini, A., Bianchini, I., Garzonio, A., & Mercuri, L. (2022). Financial Technical Indicator and Algorithmic Trading Strategy Based on Machine Learning and Alternative Data. *Risks, 10(12), 225*. https://doi.org/10.3390/risks10120225

Girsang A. S., & Stanley, N. (2023). Hybrid LSTM and GRU for Cryptocurrency Price Forecasting Based on Social Network Sentiment Analysis Using FinBERT. *IEEE Access, 11, 120530–120540.* https://doi.org/10.1109/access.2023.3324535

Kaur, G., & Sharma, A. (2023). A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of Big Data, 10*(1). https://doi.org/10.1186/s40537-022-00680-6

Karau, S. (2023). Monetary policy and Bitcoin. *Journal of International Money and Finance, 137,* 102880. https://doi.org/10.1016/j.jimonfin.2023.102880

Li, W., Jin, B., & Quan, Y. (2020). Review of Research on Text Sentiment Analysis Based on Deep Learning. *OALib, 07*(03), 1–8. https://doi.org/10.4236/oalib.1106174

Nair, M., Marie, M. I., & Abd-Elmegid, L. A. (2023). Prediction of cryptocurrency price using time series data and deep learning algorithms. *International Journal of Advanced Computer Science and Applications, 14*(8). https://dx.doi.org/10.14569/IJACSA.2023.0140837

Parente, M., **Rizzuti, L., & Trerotola, M. (2024). A profitable trading algorithm for cryptocurrencies using a Neural Network model.** ***Expert Systems with Applications, 238, 121806****–121806.* **https://doi.org/10.1016/j.eswa.2023.121806**

Plevris, V., Solorzano, G., Bakas, N., & Seghier, M. B. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. *8th European Congress on Computational Methods in Applied Sciences and Engineering*. https://doi.org/10.23967/eccomas.2022.155

Rahaman, A., Bitto, A. K., Biplob, K. B. Md. B., Bijoy, Md. H. I., Jahan, N., & Mahmud, I. (2023). Bitcoin trading indicator: a machine learning driven real time bitcoin trading indicator for the crypto market. *Bulletin of Electrical Engineering and Informatics, 12*(3), 1762–1772. https://doi.org/10.11591/eei.v12i3.4486

Winotoatmojo, H. P., Setyawan, A. A., Hendraningrat, A. R., & Setiawati, J. G. (2024). Cryptocurrency market dynamics: Analyzing trends and patterns in bitcoin. *Dinasti International Journal of Education Management And Social Science*, *5*(4), 558–564. https://doi.org/10.31933/dijemss.v5i4.2553