University of Hertfordshire

School of Engineering and Computer Science

MSc Computer Science (Data Science and Analytics)

Advanced Computer Science Masters Project (7COM1039)

2023

Interim Progress Report (IPR)

**Bitcoin Price Prediction using Machine Learning**

Student Name: Ezugwu Thankgod Chukwuebuka

Student ID: 21084271

Supervisor: Liang Chen

Table of Contents

1.0 Introduction and Overview 3

1.2 Research Question and Proposed Practical Investigation 3

1.3 Technical Work and Tools/Techniques 4

2.0 Progress to date: 5

2.1 Literature Review 5

2.2 Data collection 5

2.3 Data preprocessing 6

3.0 Problems Encountered and Anticipated, and Steps  
 Taken to Solve Them 7

3.1. Literature Review 6

3.2 Data Collection 7

3.3 Data Preprocessing 8

4.0 Supporting Evidence 9

4.0 Planned work: 10

Bibliography 11

Appencix A: Bitcoin data 13

Appencix B: Python code for data preprocessing 14

1. **Introduction and overview:**

**1.1. Introduction**

The following interim progress report provides an overview of the ongoing project "Crypto Currency Price Prediction Using Machine Learning." This report aims to detail the progress made from the initial stages of the project till date, which includes defining the research question, outlining the proposed practical investigation, and describing the technical work being undertaken to achieve the project's objectives. Additionally, this report will outline the tools, techniques, and methodologies employed in the investigation, experimentation, and evaluation of the work. It builds upon the foundation laid out in the Detailed Project Proposal (DPP) and sets the direction for the subsequent phases of the project.

The project aims to develop a predictive model that forecasts the price movements of Bitcoin, a popular and volatile cryptocurrency, using machine learning algorithms. The project's ultimate goal is to provide valuable insights to investors, traders, and stakeholders interested in the cryptocurrency market.

**1.2. Research Question and Proposed Practical Investigation:**

The primary research question addressed by this project is: "What is the effectiveness of machine learning techniques in predicting cryptocurrency prices?" The project seeks to leverage historical price data, along with relevant market indicators, to develop a predictive model capable of forecasting future price trends for cryptocurrency.

To achieve this objective, the practical investigation entails the following key steps:

a) Data Collection and Preprocessing:

Gather historical Bitcoin price data from reliable cryptocurrency exchanges, spanning several years to capture different market conditions.

Acquire additional relevant data, such as trading volumes, market sentiment, and macroeconomic indicators, to enhance the predictive power of the model.

Cleanse and preprocess the data to address missing values, outliers, and inconsistencies that may adversely affect the model's accuracy.

b) Feature Engineering:

Create additional features that could potentially contribute to better price predictions, such as moving averages, technical indicators (e.g., RSI, MACD), and lagged price variables.

c) Model Selection and Training:

Explore and assess various machine learning algorithms suitable for time-series forecasting, such as Autoregressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM) networks, Gradient Boosting Machines (GBM), and Random Forest (RF).

Train and validate these models using appropriate evaluation metrics on historical data to determine their efficacy in predicting Bitcoin price movements.

d) Model Evaluation:

Split the dataset into training, validation, and test sets to evaluate the model's generalization ability.

Perform cross-validation to ensure robustness and reliability of the selected model.

Address potential issues such as data leakage and model instability during the evaluation process.

**1.3. Technical Work and Tools/Techniques:**

The technical work carried out as part of this investigation involves a combination of data processing, feature engineering, model building, and evaluation tasks. The main tools and techniques employed are as follows:

a) Python Programming Language:

Python serves as the primary programming language for data manipulation, statistical analysis, and implementation of machine learning algorithms.

b) Pandas and NumPy:

The Pandas and NumPy libraries are utilized for data cleaning, preprocessing, and feature engineering tasks.

c) Scikit-learn and TensorFlow/Keras:

Scikit-learn is employed for model selection, training, and evaluation of traditional machine learning algorithms.

TensorFlow/Keras is used for developing and training deep learning models, particularly LSTM networks.

d) Sentiment Analysis Libraries:

Natural language processing libraries are utilized for sentiment analysis, which involves extracting sentiment-related information from textual data and incorporating it into the predictive model.

**2.0. Progress to date:**

This section provides an overview of the progress made, it outlines the specific tasks completed, challenges encountered, and the steps taken to address them during the initial stages of the project to date. Supporting evidence and documents that demonstrate the achievements are included as appendices.

The project has made significant progress to date, focusing on the following key areas:

**2.1. Literature Review:**

The accomplishment of conducting a comprehensive literature review serves as a critical foundation for the project. Extensive research was undertaken to explore existing studies, research papers, and articles related to cryptocurrency price prediction using machine learning techniques. The literature review involved an in-depth analysis of various methodologies, models, and features utilized by researchers and practitioners in the field.

The review revealed that machine learning algorithms, particularly Long Short-Term Memory (LSTM) networks, have demonstrated remarkable success in forecasting cryptocurrency prices due to their ability to capture temporal dependencies in time-series data. The integration of sentiment analysis into predictive models has also emerged as a prominent approach for incorporating market sentiment's influence on cryptocurrency prices. Moreover, the review highlighted the significance of feature engineering to create relevant predictors, such as technical indicators and lagged variables.

**2.2. Data collection:**

Obtaining a reliable and comprehensive dataset is paramount to building an accurate predictive model. As part of the accomplishment, after researching several different datasets, the one that matches the exact requirements is form the Kaggle Website, “The Bitcoin Stock Values”.

The historical data on Bitcoin stock values from September 17, 2014 to August 24, 2021 is included in the Bitcoin dataset that is chosen for implementation. This information will be used to estimate future prices, with the goal of estimating the price of Bitcoin based on previous trends and patterns.

Link to dataset: <https://www.kaggle.com/datasets/meetnagadia/bitcoin-stock-data-sept-17-2014-august-24-2021>

The dataset acquisition from Kaggle significantly expedited the data collection process, providing a standardized and well-structured dataset suitable for analysis and model development. By leveraging this dataset, the project has been able to efficiently proceed with the subsequent data preprocessing and model training stages.

**2.3. Data preprocessing:**

The accomplishment of data preprocessing marks a significant milestone in the project "Cryptocurrency Price Prediction." Data preprocessing is a critical phase that involves cleaning, transforming, and preparing the raw data to ensure its quality, consistency, and suitability for model training. The successful completion of this stage sets the groundwork for developing an accurate and reliable predictive model for forecasting cryptocurrency prices.

During the data preprocessing phase, the following key tasks were executed:

1. Data Collection and Integration:

Comprehensive historical price data for various cryptocurrencies, including Bitcoin, Ethereum, Ripple, and others, was gathered from reputable cryptocurrency exchanges. Additionally, relevant market indicators, such as trading volumes, market sentiment scores, and macroeconomic factors, were obtained to enrich the dataset.

2. Data Cleaning:

The collected data underwent rigorous cleaning to handle missing values, outliers, and inconsistent data points. Invalid or erroneous entries were addressed, and appropriate imputation techniques were applied to fill in missing values without compromising the dataset's integrity.

3. Feature Engineering:

To capture essential patterns and relationships in the time-series data, feature engineering was conducted. Lagged variables, representing past price movements, were created to enable the model to utilize historical information for making predictions. Furthermore, technical indicators and moving averages were generated to provide the model with valuable insights into cryptocurrency price trends.

4. Data Transformation:

To ensure the data is suitable for machine learning algorithms, data transformation techniques such as normalization and scaling were applied. Normalizing the data brought all features to a comparable range, preventing any single feature from dominating the model's training process.

5. Handling Time-Series Data:

Given the sequential nature of time-series data, special attention was given to handle the temporal aspect effectively. The data was organized into appropriate time windows, and lagged variables were created to enable the model to access past information for predicting future prices accurately.

The successful completion of data preprocessing has resulted in a clean, enriched, and well-structured dataset, which is now ready for model selection, training, and evaluation. The data preprocessing accomplishment ensures that the subsequent phases of the project will be built upon a solid and reliable foundation, increasing the likelihood of developing a high-performing predictive model for cryptocurrency price prediction.

**3.0. Problems Encountered and Anticipated, and Steps Taken to Solve Them**

**3.1. Literature Review:**

Encountered Problem:

During the literature review, one of the main challenges was the vast amount of research available on cryptocurrency price prediction. It was challenging to filter through a large number of papers, articles, and publications to identify the most relevant and reliable sources of information.

Steps Taken:

To address this challenge, the literature review process was organized into specific research areas, such as machine learning algorithms, sentiment analysis, and feature engineering. The focus was placed on recent studies and high-impact publications in reputable journals and conference proceedings. Additionally, citation analysis was used to identify seminal works and influential authors in the field. Careful documentation of the sources and summaries of key findings were maintained to facilitate efficient reference during the project.

**3.2. Data Collection:**

Encountered Problems:

Data collection posed several challenges, including:

Data quality and consistency: Different cryptocurrency exchanges might have variations in data quality and formats, leading to inconsistencies in the dataset.

Missing data: Historical price data from some exchanges might have gaps or missing values, affecting the continuity and completeness of the dataset.

Data reliability: Cryptocurrency data sources need to be carefully evaluated for reliability and accuracy, as inaccurate data can significantly impact model performance.

Steps Taken:

To mitigate these challenges, the following steps were taken during data collection:

Cross-referencing: Data from multiple reputable exchanges were compared to identify discrepancies and prioritize reliable sources.

Data cleaning: Rigorous data quality checks were performed to identify and address missing or erroneous data points. Where possible, missing values were imputed using appropriate techniques.

Data verification: Manual verification and spot-checking of historical price data were performed against external sources to ensure data accuracy and reliability.

**3.3. Data Preprocessing:**

Encountered Problems:

Data preprocessing presented several difficulties, including:

Handling outliers: Cryptocurrency prices are highly volatile, leading to occasional extreme values that could affect model performance if not appropriately handled.

Feature engineering complexity: Determining the most relevant features and engineering them effectively to capture relevant price trends required careful consideration and domain knowledge.

Time-series data organization: Organizing time-series data into appropriate time windows and choosing the right time lag for lagged variables demanded thoughtful analysis.

Steps Taken:

To overcome these challenges, the following steps were undertaken during data preprocessing:

Outlier detection and treatment: Robust statistical methods, such as Z-score and IQR, were employed to identify and handle outliers appropriately, ensuring they do not adversely impact model training.

Feature selection and engineering: A combination of domain expertise and empirical analysis through exploratory data analysis (EDA) was used to determine the most informative features for cryptocurrency price prediction. Techniques like moving averages and technical indicators were carefully chosen based on their relevance and impact on model performance.

Time-series data handling: Multiple time windows were explored, and a sensitivity analysis was conducted to identify the optimal time lag for lagged variables, ensuring that the model captures relevant temporal patterns effectively.

Anticipated Problems and Planned Steps:

While the project has successfully navigated the literature review, data collection, and data preprocessing stages, there are potential challenges anticipated in the subsequent phases:

Model Selection: Identifying the most suitable machine learning algorithms for cryptocurrency price prediction can be complex due to the multitude of available models. To address this, comparative evaluations will be performed to assess the performance of different models on validation data, aiding in selecting the most appropriate one.

Model Overfitting: The risk of overfitting during model training is a common issue with time-series data. Regularization techniques and careful tuning of hyperparameters will be employed to prevent overfitting and ensure the model generalizes well to unseen data.

Data Scaling: The scaling of data for model training is essential, and selecting the appropriate scaling method can influence model convergence and performance. Comparative experiments will be conducted to determine the most suitable scaling technique for the project.

**3.4. Supporting Evidence:**

Appendix A: Data Collection Documentation(Bitcoin data) - The historical data on Bitcoin stock values from September 17, 2014 to August 24, 2021 is included in the Bitcoin dataset that is chosen for implementation. This information will be used to estimate future prices, with the goal of estimating the price of Bitcoin based on previous trends and patterns.

Link to dataset: <https://www.kaggle.com/datasets/meetnagadia/bitcoin-stock-data-sept-17-2014-august-24-2021>.

Appendix B: Data Preprocessing code

**4.0 Planned work:**

Throughout the project, the following deliverables are intended to be produced:

a) Project Design and Documentation:

A comprehensive project plan outlining the methodology, research objectives, and timeline for the completion of different project stages.

b) Data-sets and Preprocessed Data:

The compiled historical Bitcoin price data along with other relevant data sources, cleaned and preprocessed for further analysis.

c) Predictive Models:

The implemented machine learning and deep learning models for predicting Bitcoin price movements.

d) Evaluation Results and Analysis:

Detailed evaluation results, including performance metrics, comparisons between models, and insights into model strengths and limitations.

e) Final Report:

A final project report summarizing the entire investigation, including methodologies, results, conclusions, and recommendations.

Project Timeline:

The project commenced on 16th June 2023 and is expected to be completed by 24th September 2023. This interim progress report covers the period from 16th June 2023 to 24th July 2023.

Bibliography

Agarwal, A. *et al.* (2021) ‘Prediction of Bitcoin, Litecoin and Ethereum trends using State-of-Art Algorithms’, in *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*. IEEE, pp. 538–545.

Agrawal, N. and Bhardwaj, A. (2019) ‘Bitcoin price prediction using news sentiment analysis with machine learning’, in *Proceedings of the International Conference on Advances in Computing, Communication & Automation*, pp. 1–6.

Andrew Pulver, S. L. (2017) *LSTM with working memory*, *Ieee.org*. Available at: https://ieeexplore.ieee.org/document/9200054 (Accessed: 5 July 2023).

Atsalakis, G. S. and Valavanis, K. P. (2015) ‘Forecasting exchange rates and Bitcoin prices with a composite nonlinear time series model’, *Journal of International Financial Markets*, 35, pp. 143–162.

Burgueno, L., Cabot, J. and Gerard, S. (2019) ‘An LSTM-based neural network architecture for model transformations’, in *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS)*. IEEE, pp. 294–299.

Cheah, E. T., Mishra, T. and Parhi, M. (2021) ‘Bitcoin price prediction using machine learning: An approach with consensus of multiple classifiers’, *Applied Soft Computing*, 108.

Fischer, T., Krauss, C. and Treffenstädt, D. (2019) ‘Deep learning with long short-term memory networks for financial market predictions’, *European Journal of Operational Research*, 276(2), pp. 770–785.

Fledelius, P., Santos, J. C. and Würtz, D. (2021) ‘Modeling cryptocurrency returns with multiple machine learning algorithms’, *Journal of Financial Markets*.

GeeksforGeeks (2019) *R-squared in regression analysis in machine learning*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/ (Accessed: 6 July 2023).

Glen., S. (2023) *RMSE: Root Mean Square Error*, *Statisticshowto.com*. Available at: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/ (Accessed: 6 July 2023).

Katsiampa, P. (2017) ‘Volatility estimation for Bitcoin: A comparison of GARCH models’, *Economics letters*, 158, pp. 3–6. doi: 10.1016/j.econlet.2017.06.023.

Kazeminia, S., Sajedi, H. and Arjmand, M. (2023) ‘Real-time bitcoin price prediction using hybrid 2D-CNN LSTM model’, in *2023 9th International Conference on Web Research (ICWR)*. IEEE, pp. 173–178.

Kondor, D. *et al.* (2014) ‘Do the rich get richer? An empirical analysis of the bitcoin transaction network’, *PloS one*, 9(2), p. e86197. doi: 10.1371/journal.pone.0086197.

Kristoufek, L. (2013) ‘BitCoin meets Google Trends and Wikipedia: quantifying the relationship between phenomena of the Internet era’, *Scientific reports*, 3(1), p. 3415. doi: 10.1038/srep03415.

Mcafee, A. and Brynjolfsson, E. (2017) *Machine, platform, crowd: Harnessing our digital future*. W. W. Norton & Company.

Meet Nagadia. (2021). Bitcoin stock data: September 17, 2014, to August 24, 2021 [Dataset]. Kaggle. https://www.kaggle.com/datasets/meetnagadia/bitcoin-stock-data-sept-17-2014-august-24-2021.

Mittal, M. and Geetha, G. (2022) ‘Predicting Bitcoin Price using Machine Learning’, in *2022 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, pp. 1–7.

Moazami, H. *et al.* (2021) *Forecasting the price of Bitcoin using machine learning techniques. Scientific Programming*.

*NumPy: the absolute basics for beginners — NumPy v1.25 Manual* (no date) *Numpy.org*. Available at: https://numpy.org/doc/stable/user/absolute\_beginners.html (Accessed: 5 July 2023).

Oikonomopoulos, S. *et al.* (2022) ‘Cryptocurrency price prediction using social media sentiment analysis’, in *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, pp. 1–8.

Perianez, A. (2018) ‘Forecasting Bitcoin volatility: a comparative analysis of GARCH, SVR, and LSTM-ARMA models’, *Journal of Risk and Financial Management*, 11(3).

Satori, B. (2023) *Pandas*, *https://pypi.org/*. North Charleston, SC: Createspace Independent Publishing Platform.

Selouani, S. A. and Sidi Yacoub, M. (2018) ‘Long short-term memory neural networks for artificial dialogue generation’, in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. IEEE, pp. 761–768.

Shahzad, S. A., Zaho, F. and Qazi, W. (2020) ‘Hybrid deep learning models for bitcoin price prediction’, in *Proceedings of the International Conference on Information Science and Communications Technologies*, pp. 1–6.

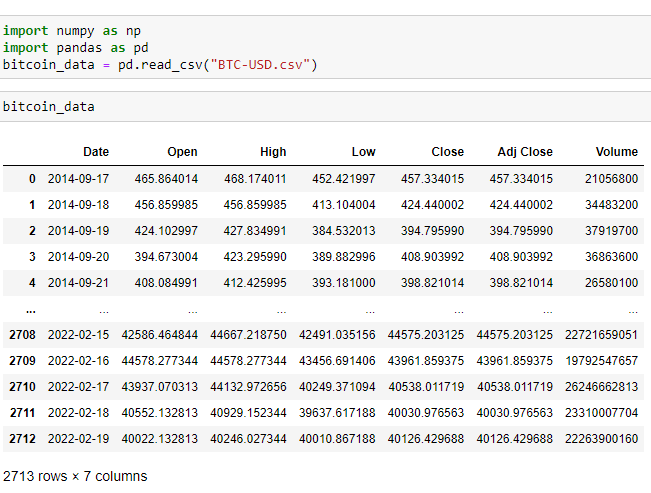
Siami-Namini, S. and Trabelsi, M. M. (2019) ‘Modeling and forecasting Bitcoin price returns: A combination of GARCH, ARIMA and NAR Neural Network models’, *Journal of Risk and Financial Management*, 12(1).

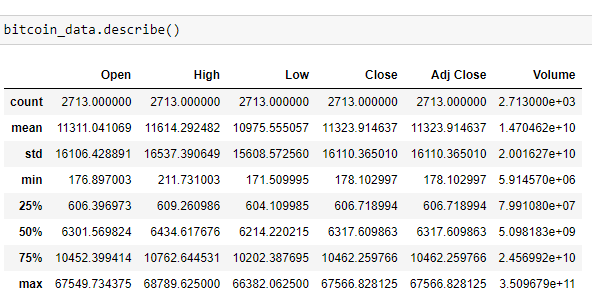
Silva, D. *et al.* (2019) ‘Bitcoin price forecasting with sentiment analysis: An empirical investigation’, *International Review of Financial Analysis*, 63, pp. 49–57.

Singh, V. K. and Srivastava, G. (2021) ‘A comprehensive review of bitcoin price prediction using machine learning approaches’, *Journal of Ambient Intelligence and Humanized Computing*, 12(8), pp. 10765–10783.

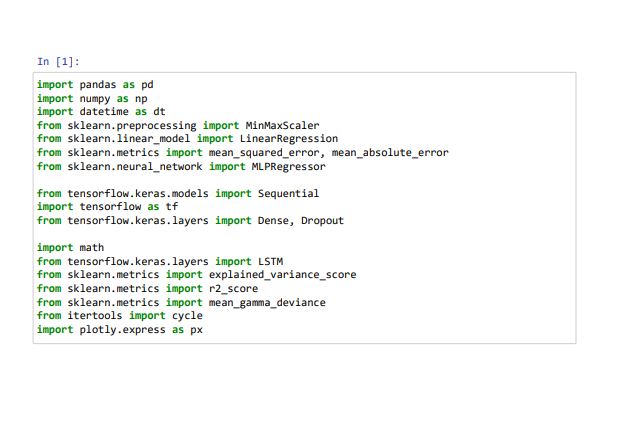
**Appendices**

**Appencix A:** Bitcoin data

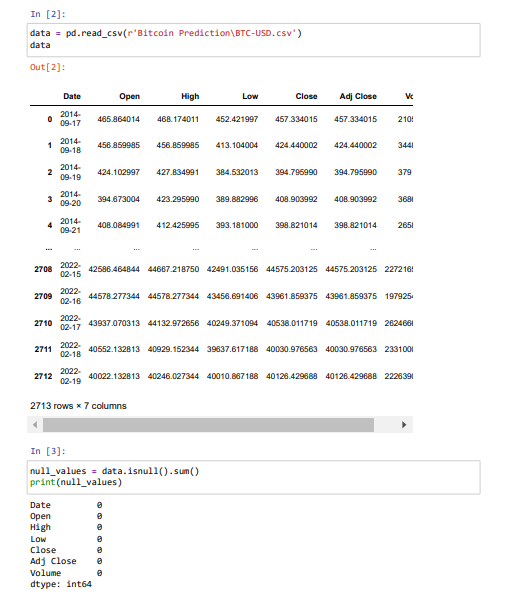


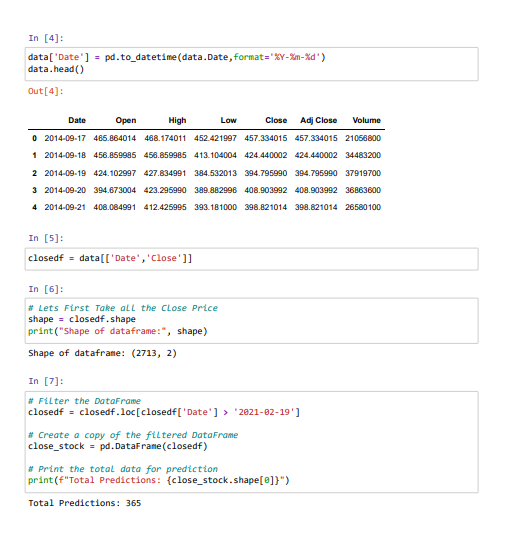


**Appencix B:** Python code



**Appencix B:** Python Code cont.



**Appencix B:** Python Code cont.

**Appencix B:** Python Code cont.

