



Crypto Currency Market Price Prediction Using Data Science Technique

R. Merline Snowfia¹, Joy Suganthi Bai¹

¹Department of Computer science and Engineering, Grace College of Engineering, Tamil Nadu, India.

ABSTRACT

Cryptocurrency, also referred to as digital currency, protects transaction records, controls the creation of new coins, and verifies ownership transfers by employing high quality encrypting. A ledger that exists as a computerised database maintains records of coin ownership. It is challenging to estimate a price of the crypto currency because it is used on a broad scale now and its share can increase or drop suddenly. For the purpose of predicting Cryptocurrency prices, a machine learning model is developed in this paper. To build a more accurate model for outcome prediction, data science technique is applied. The key steps in creating a successful model are variable identification and data comprehension. The performance of each machine learning algorithm is evaluated by comparing its reliability to the precision of the pre-processed data. Improved performance measures for the model, such as accuracy, recollection, and scoring, are also taken into account for evaluation. To forecast the cryptocurrency, one employs a machine learning system.

Key words: Bitcoin, Ethereum, Cardamom (ADA), Binance Coin (BNB)

1 Introduction

1.1 Data Science

Data science is a multidisciplinary field that integrates analytical methods, approaches, procedures and technologies to obtain knowledge and skill insights from both organized and unstructured data and apply them to a wide range of application domains.

Peter Naur coined the phrase "data science" in 1974 as a substitute for the word "computer science." In 1996, the International Federation of Classification Societies became the first organisation to implement this.

Particularly emphasise the topic of data science. But there was still disagreement over the term. The phrase "data science" was initially coined in 2008 by the foundation heads of data and analytics operations at Facebook and LinkedIn, D.J. Patil and Jeff Hammerbacher. It has quickly risen to the top of the most popular and sought-after careers.

1.2 Data Scientist

To extract meaningful insights from data, the field of study known as "data science" brings together. Combining math, business savvy, tools, algorithms, and machine learning techniques



can be summed up as data science. These elements all interact to enable us to unearth hidden patterns or insights in raw data, which can be quite helpful when making crucial business decisions.

Data scientists examine at the questions that need to be answered as well as the best places to look for the relevant data. Along with the capacity to gather, clean, and display data, they also possess business sense and analytical abilities. Data scientists are employed by businesses to gather, handle and analyses significant amounts of unstructured data.

1.3 Important Skills for a Data Scientist

- **Programming:** SQL, Python.
- **Machine Learning:** Natural Language Processing, Classification, Clustering.
- **Data Visualization:** Tableau, SAS, D3.js, Python, Java, R libraries.
- **Big data platforms:** Mongo DB, Cloud era.

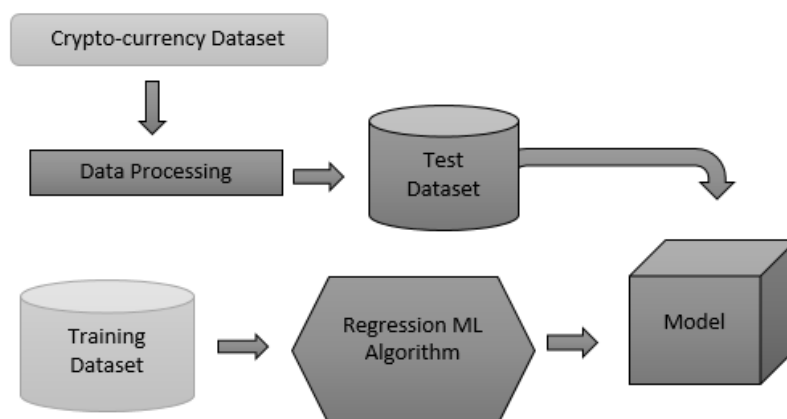
1.4 Existing System

They provided a computer method for locating and classifying social media-based crypto currency pump and dump operations. The method used was able to predict, with reasonable certainty, whether or not an attack on that coin was currently occurring on Telegram and whether the eventual pump procedure would be successful in attaining the expected target prices. They had used financial and Twitter data regarding a specific coin. Additionally, they examined the actions of users who participated in pumping operations and saw that tweets about crypto currencies were frequently sent by Twitter bots in the immediate aftermath of the attack. Scammers frequently used Telegram to plan and coordinate pump and dump schemes.

1.5 Proposed System

Cryptocurrency operates uniquely, it might be challenging to foresee potential outcomes. The suggested framework involves creating a model that can forecast the price. Data science processes are applied throughout the steps of the developed framework, from variable characterization to model creation. The first step in the procedure is to identify the variables, such as research variables, in the target column.

Pre-processed data is then utilized to build a model by separating the dataset into a 7:3 ratio, where 70% of the data is used for training, during which the model identifies sequencing, and the remaining testing data is used to verify the usefulness of the data. The price of the coin can also be predicted using the regression model.



1.6 Advantages

1. For improved prediction, the performance metrics and machine learning algorithms are compared.
2. Business predictions may be made with a high degree of accuracy to machine learning models.

2 Algorithm Techniques

2.1 Algorithm Explanation

A computer software collects knowledge from the data input given to it and uses that knowledge to classify subsequent observations. This process is known as classification and is used in statistics and machine learning. It's possible for this data set to just contain two classes (for example, indicating whether the individual is male or female or if the message is spam or not). Speech recognition, handwriting recognition, biometric identity, document classification, etc. are a few instances of classification issues. Algorithms are taught by supervised learning using labelled data. Based on patterns and associations with the previously unlabeled new data, the algorithm determines which label new data should receive after understanding the data.

2.2 Used Python Packages

sklearn

- A number of ML methods are included in the Python machine learning package called Sklearn.
- In this situation, several of its components, such as accuracy score, train test split and decision tree classifier.

NumPy

- It is a Python module for numerical computations that offers quick math functions.
- It is utilized for data manipulation and reading from numpy arrays.

Pandas



- Can be used to read and write many file types.
- Data frames make it simple to manipulate data.

Matplotlib

- Finding trends in a particular dataset can be aided through data visualization.
- Data frames make it simple to manipulate data.

3 Software Description

Anaconda, a free and open source Python and R implementation for scientific computing, aims to streamline management system to ensure and deployments. Versioning of packages is managed via the "Conda" package management system. More than 12 million people use the Anaconda distribution, which includes more than 1400 well liked data science applications accessible with Windows, Linux and MacOS. Many of the features offered by conda packages are also found in pip packages, and in most situations they can coexist. The programme can be used to generate unique packages that can then be posted to Anaconda Cloud, PyPI or other. Python 2.7 and Python 3.7 are both installed by default with Anaconda2 and Anaconda3, respectively.

3.1 Data Pre-processing

The machine learning (ML) model's error rate is determined through validation procedures, and it is believed to be as close as feasible to the dataset's real error rate. It is not require the validation approaches if the volume of the data is sufficient to be representative of the population. However, working with data samples that could not be a realistic reflection of the population of a given dataset in real-world circumstances. Locating duplicate values, missing information, and information regarding the data type, such as whether it is a float variable or an integers are all important. Subset of data used to tune model hyper parameters while providing an objective assessment of model fit on the training dataset. Several of these include some minor inaccuracies. In some cases, data loss can have a more serious cause. It's critical from a statistical perspective to comprehend these various missing data kinds. The kind of missing data will affect how it is filled in, how it is detected, how it is handled statistically, and how it is handled using both basic imputation and detailed statistical methods. It's critical to comprehend sources of missing data before incorporating them into the programming. Here are a few typical explanations for missing data.

3.2 Exploration data analysis of visualization

A crucial skill in applied statistics and machine learning is data visualization. In actuality, the primary emphasis of statistics is on quantitative estimates and data descriptions. Data visualization offers a crucial set of tools for gaining a qualitative insight. When examining and learning about a dataset, this can be helpful for finding trends, outliers, corrupt data and much more. Data visualizations can be used to explain and depict key relationships in plots and charts in a way that is more visceral and engaging to stakeholders than relying just on subject specific



connections or relevance judgments. It will suggest a deeper look at some books suggested at conclusion because data visualization and exploratory data analysis are entire fields in themselves. The sample below examines distinct algorithms,

- Logistic Regression
- Decision Tree Regression,
- Random Forest Regression,
- Lasso Regression and Support Vector Regression

3.3 Logistic Regression

It is a scientific technique for analyzing a set of data in which one or more independent variables affect the result. The objective of logistic regression is to identify the model that best describes the relationship between a set of independent (predictor or explanatory) factors.

To ascertain the probability of a group of variables, multiple regressions, a machine learning classification technique, are utilized. With binary data coded as 1 (yes, success, etc.) or 0 (no), the dependent variable in regression models is a binary variable (no, failure, etc.)

The logistic regression model, in other words, predicts $P(Y=1)$ as an assumptions for logistic regression.

3.4 Decision Tree Regression

In classification models or as a tree structure, regression decision trees are used. It builds a decision tree iteratively as a dataset is divided into progressively smaller subsets. The result is a tree with leaf nodes and decision nodes. A decision tree generates classification or regression models that resemble trees.

3.5 Random Forest Regression

In a monitored classification model called Random Forest Regression, stagnation is treated with the quartet learning strategy. Using a learning technique, projections from various machine learning algorithms are combined to produce predictions that are more accurate than those from a single model. After many classification decision trees have been fitted to various subsamples of the dataset, a random forest is a nonlinear estimator that uses averaging to improve projected accuracy and decrease over fitting. The max samples argument determines the size of the sub sample if Bootstrap = True (the default), else each tree is built using the complete dataset.

3.6 Support Vector Regression

For classification and regression issues, Support Vector Machines (SVM) are often and extensively utilized in learning algorithms. A refinement of the classification problem, the extrapolation problem has the model returning a constant voltage rather than an output from a limited collection. In other terms, a recurrence model estimates a multivariable statistic with continuous values.



3.7 Lasso Regression

A regularization method is lasso regression. For a forecast that is more accurate, regression approaches are preferred. This model benefits from shrinking. When data values converge on the mean, this process is known as shrinkage. Simple, sparse models are encouraged by the lasso technique. A regularization method is lasso regression. It is used in place of regression-based methods to obtain more precise predictions. This model benefits from shrinking. When data values converge on the mean, this process is known as shrinkage.

4 Categorization

4.1 Flask (Framework)

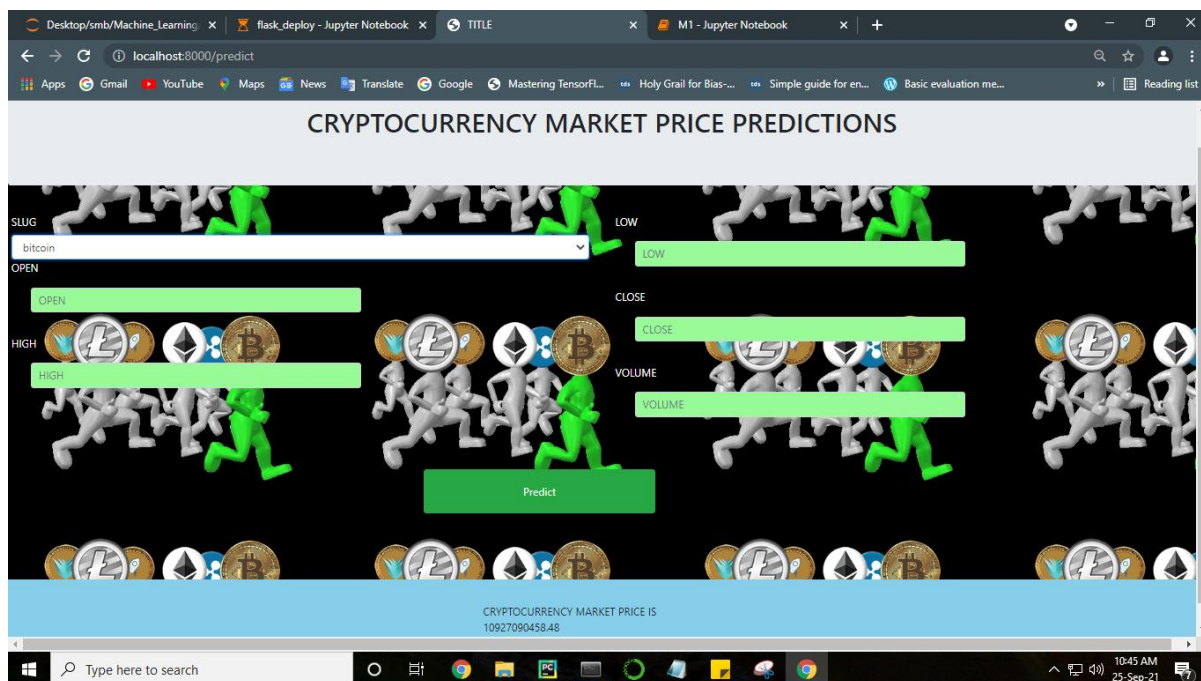
Flask is a web framework for micro systems with a Python foundation. It is categorized as a micro-framework because no specific tools or libraries are needed. It is free of any components that a database abstraction layer or form validation. Nevertheless, the flask framework has extensions that can be used to add application functionalities in a similar way to how they were incorporated into the core framework. In addition to utilities connected to well-known frameworks, there are extensions for object relational mappers, form validation, upload handling, several open authentication methods.

5 Conclusion

Data processing and handling, non - informative analysis, eventually determine, model construction and evaluation came first in the analysis phase. The highest accuracy score on the public test data set will be determined. This programme can assist in determining the rate of the crypto currency market.



Article Title: Crypto Currency Market Price Prediction Using Data Science Technique



Reference

- [1] R. Beck; M. Avital; M. Rossi; J.B. Thatcher, Year: 2017, "Block chain technology in business and information systems research", Business & Information System Engineering, Vol: 59, no:6, pp. 381-384.
- [2] S. Gu; B. Kelly; D. Xiu, Year: 2020, "Empirical asset pricing via machine learning". The Review of Financial Studies, Empirical asset pricing via machine learning Vol: 33, no:5, pp. 2223-2273.
- [3] Feng; Guan hao; Stefano Giglio; Dacheng Xiu, Year: 2020, "Taming the factor zoo: A test of new factors." The Journal of Finance 75, no. 3, pp. 1327-1370.
- [4] P. Jaquart; D. Dann; C. Martin, Year: 2020, "Machine learning for bitcoin pricing - a structured literature review". WI 2020 Proceedings, GITO Verlag pp. 174-188.
- [5] E.F. Fama, Year: 1970, "Efficient capital markets: a review of theory and empirical work", J Finance, 25, pp. 383-417.
- [6] A.W. Lo, Year: 2019, "The adaptive markets hypothesis Adaptive Markets", Princeton University Press, vol. 30, pp. 176-221.
- [7] T. Fischer; C. Krauss, Year: 2018, "Deep learning with long short-term memory networks for financial market predictions", Eur J Oper Res, 270, pp. 654-669.
- [8] B. Krollner; B. Vanstone; G. Finnie, Year: 2010, "Financial time series forecasting with machine learning techniques: a survey Proceedings of European Symposium on Artificial Neural Networks: Computational and Machine Learning", Springer, pp. 1-7.
- [9] G.W. Schwert, Year: 2003, "Anomalies and market efficiency Handbook of the Economics of Finance", Handbook of the Economics of Finance, Elsevier, pp. 939-974.
- [10] A.H. Dyhrberg, Year: 2016, "Bitcoin gold and the dollar - a GARCH volatility analysis", Finance Res Lett, 16, pp. 85-92.