# *CRYPTO-CURRENCY MARKET CAP PREDICTION USING MACHINE LEARNING*

Hemkiran S
*Department of Computer Science and Engineering*
*PSG Institute of Technology and Applied Research*
Coimbatore, India
hemkiran@psgitech.ac.in

Rajiv Sunduram G
*Department of Computer Science and Engineering*
*PSG Institute of Technology and Applied Research*
Coimbatore, India
rajivsunduram2001@gmail.com

Shyam Ganesh T
*Department of Computer Science and Engineering*
*PSG Institute of Technology and Applied Research)*
Coimbatore, India
shyamganesh2307@gmail.com

***Abstract*— Crypto-currency is considered as the digital asset designed to work as a medium of exchange, whereas individual coin ownership records are stored in computer databases using strong cryptography techniques to secure transactional records. The value of crypto-currencies is determined by the basic demand-supply chain in the market. If the supply is limited, then the demand increases which leads to the increase in the price of the crypto-currencies. In this paper, data of about 887 crypto-currencies was utilized for evaluating the machine learning algorithms. The Market Capitalization of the crypto-currencies can be predicted using suitable parameters. Since trading crypto-currencies is considered as the source of income for many investors, this system aids them in making suitable right decisions. This project compares various regression techniques of machine learning algorithms and suggests the best model based on its performance metrics.**

***Keywords*— Crypto currency, market cap, normalization, prediction system, regression.***

## I. INTRODUCTION

The deficit of proper understanding of data is the significant reason for the failure of many companies in their investment initiatives. The process of getting useful insights from the data is a hardened deal. The data science and machine learning concepts enable companies to efficiently understand gigantic data from multiple sources and derive valuable insights to make smarter data-driven decisions. A cryptocurrency is a digital or virtual currency that is secured by cryptography, which makes it nearly impossible to forged. Many cryptocurrencies are decentralized networks based on block chain technology, which is a distributed ledger enforced by a disparate network of computers.

Many companies and individuals these days are currently investing in the crypto-currency market. Without prior knowledge of future value of coins, many investors might take wrong decisions. Therefore, to avoid this circumstances, here the work is done to predict the market cap of crypto-currencies beforehand using the major factors that are influencing the crypto market. Investors can use market capitalization to assess the value of a crypto currency they are considering to buy. Market capitalization is a key measure of profitability that is also used in equations to determine price-to-earnings and other significant metrics. The investment community uses this measure to determine a company's size and basically how these crypto market is valuing the company. The predictive analysis of this data will help the investors to diversify their portfolio. This work may help the investors in making right decisions in their investment strategies.

## II. LITERATURE SURVEY

There are many existing work on the field of cryptocurrency as digitization is made popular nowadays. A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, and A. Baronchelli proposed in their work that the popularity of cryptocurrencies was raised due to the exponential growth of their market capitalization, which peaked at more than $800 billion [1]. Based on the recent study made by G. Hileman and M. Rauchs, there were more actively traded cryptocurrencies between millions of private as well as institutional investors from different transaction networks and access to the market has become easier over time [2]. S. Nakamoto suggested that the stock market is diverse and provides investors with variety of products and Bitcoin was precisely designed as a medium of exchange [3]. J. Barrdear and M. Kumhof stated that the bitcoins provide more control and security for personal and financial information as mentioned in [4].

S. M. Raju and Ali Mohammad Tarif suggested Twitter and Reddit posts to analyse the public sentiment of people to predict the value of bitcoin. They implemented recurrent neural networks (RNN) along with long short-term memory cells (LSTM) to achieve better efficiency [5]. In [6], D. Shah and K. Zhang used time series datasets along with random forest technique to predict the price of bitcoin. The research work exhibited by Ruchi Mittal, Shefali Arora and M P S Bhatia in [7], utilized multivariate linear regression for predicting highest and lowest price of cryptocurrencies by using features like open, low and close. In [8], Sean Mc Nally, Jason Roche, Simen Caton implemented certain machine learning techniques for the price prediction of bitcoin in the market.

## III. DATASETS AND FEATURES

The dataset used in the proposed work is based on Kaggle's cryptocurrencies historical data. This dataset has 6,32,218 observations and 8 attributes. To have a better understanding of the features available in the dataset, the purpose of the attributes is mentioned below. After knowing about all these features, it is decided to set the target variable as Market Cap and the remaining attributes are mentioned as shown below.

- Date – In YYYY/MM/DD format
- Open – The initial price of the coin in the start of that day.

- High – The maximum price of the coin on that particular day.
- Low – The minimum price of the coin on that particular day.
- Close – The price of the coin at the end of the particular day.
- Volume – Total value of coins that has been traded on that particular day.
- Symbols - An entity to uniquely identify a cryptocurrency.
- Market cap – Value used to define the worth of a cryptocurrency.

## IV. Methodologies

Fig. 1 represents the workflow of crypto-currency analysis, and the steps to be taken to predict the Market cap of crypto-currencies.

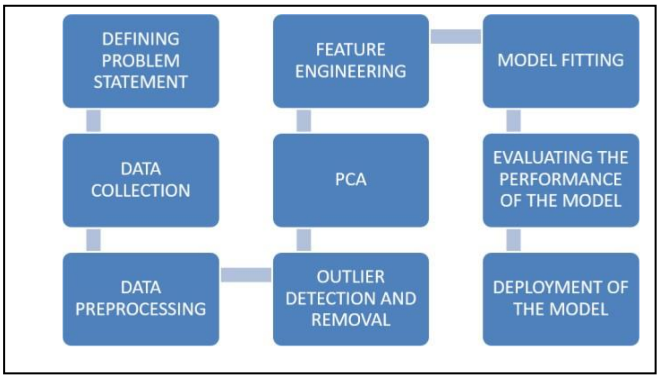
DEFINING PROBLEM STATEMENT
FEATURE ENGINEERING
MODEL FITTING
DATA COLLECTION
PCA
EVALUATING THE PERFORMANCE OF THE MODEL
DATA PREPROCESSING
OUTLIER DETECTION AND REMOVAL
DEPLOYMENT OF THE MODEL

Fig. 1. Workflow of crypto-currency analysis

The cryptocurrency analysis comprised of the following steps, such as,

*1)* Defining the problem statement and identifying the right data required to solve the problem.

*2)* Required historical data is retrieved from Kaggle website.

*3)* The attributes are clearly studied and followed by pre-processing steps.

*4)* Data Pre-processing:

a. *Statistical details of the attributes is calculated.*

b. *Missing values are found in volume and market cap and they are cleaned by removing the rows which had missing values in it.*

*5)* Outliers are detected for all attributes and detached from the dataset for further processing.

*6)* For performing the symbol feature first, the symbols are modified into categories and frequency encoding is performed on the categorical data of symbol feature.

*7)* The data is normalized using min-max scalar technique.

*8)* The independent features in the dataset is visualized against the dependent variable to find their trends like linearity, correlation etc., Fig. 2 represents the correlation matrix for the data considered here.

*9)* Correlation matrix is calculated for all the independent variables and visualized using correlation plot to find the correlation between them.

*10)* The dataset is split into two parts namely train and test splits and their corresponding split ratios are 70:30 respectively.

*11)* The independent and dependent variables are separated where Market Cap is the target variable and rest are input variables.

*12)* Models like KNN, Artificial Neural Network, Random forest regressor, XGBoost, Lasso and linear regression, are used to train the training data and their performance is evaluated on the development data.

*13)* The performance of the model was considered better in Random Forest regressor. Its performance is evaluated on the development and test sets based on the metrics such as R-square value and mean squared error.

## V. Proposed Model

### *A. Linear Regression*

Linear Regression is a machine learning algorithm based on supervised learning and performs regression task on the dataset. Linear regression performs the task to predict a dependent variable value (Y) based on a given independent variable (X). So, this regression technique finds out a linear relationship.

$$Y = aX + b \qquad (1)$$

Where, X is the explanatory variable, Y is the dependent variable, the slope of the line is a and b is the intercept.

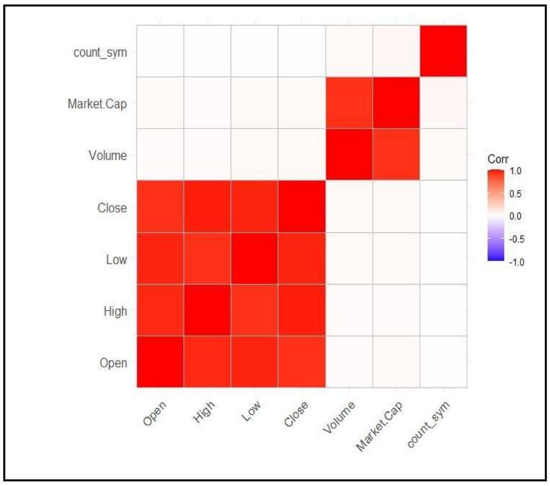

Fig. 2. Correlation matrix

### *B. Random Forest regressor*

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. Here, row and feature sampling was randomly performed to generate sample datasets for every model which is termed as bootstrap.

### C. *KNN Regression*

Another simple implementation of KNN regression is to compute the average of the numerical target of K nearest neighbours. It determines the feature similarity to perceive the new data points and this value is allotted based on the closeness of this point with its neighbours.

### D. *XGBoost*

XGBoost or extreme gradient boosting is one of the popular gradient boosting techniques having enhanced performance and speed in tree-based machine learning algorithms. It is considered to be one of the commonly used ensemble learning algorithm. This algorithm is preferred for its accuracy and faster response nature.

### E. *Lasso Regression*

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are reduced towards a central point, like the mean. The lasso procedure encourages simple, sparse models with fewer parameters. This particular type of regression is well-suited for models showing high levels of multicollinearity or to automate certain parts of model selection, like variable selection/parameter elimination.

### F. *Artificial Neural Network*

An artificial neural network (ANN) is the piece of a computing system designed to simulate the way the human brain analyses and processes information. It is the foundation of Artificial Intelligence (AI) and solves problems that would prove impossible or difficult by human or statistical standards. ANNs have self- learning capabilities that enable them to produce better results as more data becomes available.

## VI. RESULTS AND DISCUSSION

As the proposed model is a regression type of problem, the R-square and mean squared error was considered as evaluation metrics. The R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. The Mean squared error (MSE) provides the information of how a regression line is close to the set of points. It is calculated by the distances from the points to the regression line and computing its square value. Here is the table which includes the comparison of different models on the test data for the performance metrics such as R squared value and Mean Squared Error.

TABLE I. PERFORMANCE METRICS COMPARISON

The images of the graphs plotted for true market cap value against the predicted market cap value for the various models mentioned in the proposed work are attached here.

| **Model Name** | **Performance Metrics** | |
|---|---|---|
| | ***R squared value*** | ***Mean Squared Error*** |
| Linear Regression | 0.8698 | 1.8148 |
| Random Forest Regressor | 0.9956 | 6.0859 |
| KNN Regression | 0.8517 | 0.8517 |
| XGBoost | 0.9938 | 8.5878 |
| Lasso Regression | 0.8698 | 1.8148 |
| Artificial Neural Network | 0.8710 | 2.1477 |

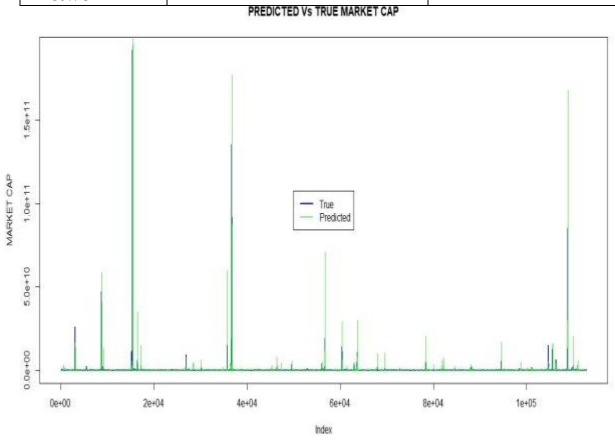

Fig. 3. Linear regression

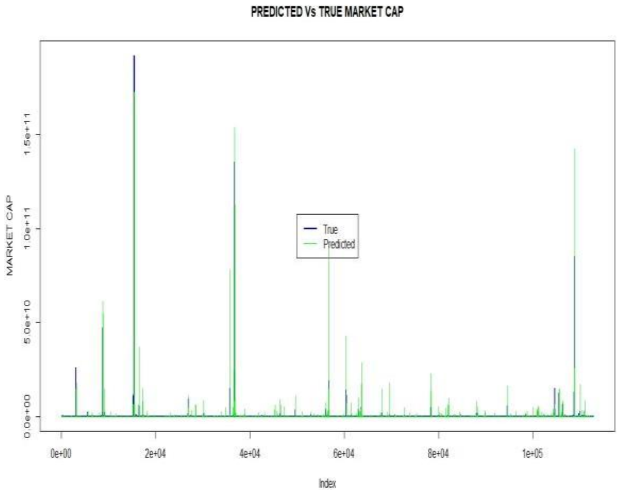

Fig. 4. Random Forest Regressor

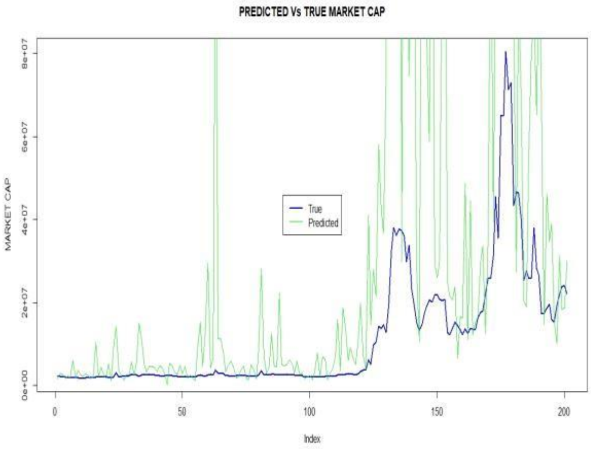

Fig. 5. KNN Regression

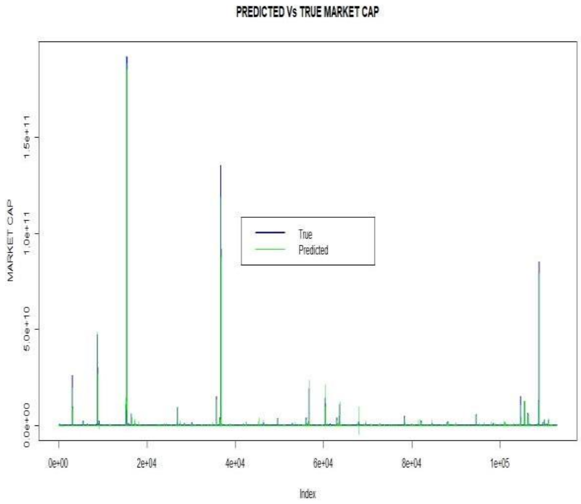

Fig. 6. XGB Regression

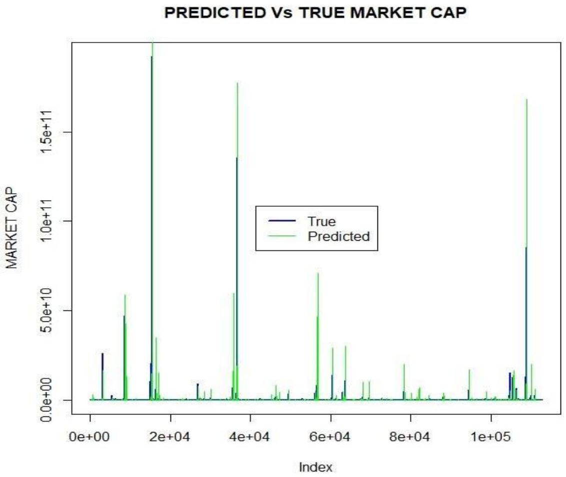

Fig. 7. Lasso Regression

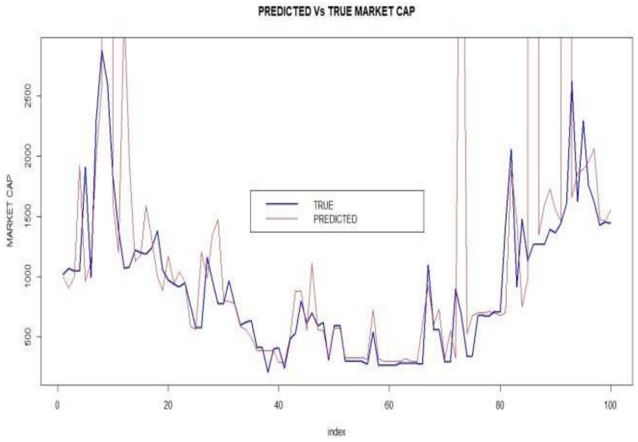

Fig. 8. Artificial Neural Networks

## VII. CONCLUSION AND FUTURE WORK

In this paper, the historical data of crypto-currencies was analysed and the Market Cap was predicted. Here, the Random Forest Regressor performs comparatively better than other algorithms in terms of R squared value. This system tries to predict the worth of the digital asset of the people who were investing in the stock market. In the future, the model may be deployed in the real world applications in a more user-friendly manner, so that user with no prior knowledge about crypto currencies can able to easily understand the processes and take better decisions in their investment strategies.

## REFERENCES

[1] A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor- Satorras, and A. Baronchelli. “Evolutionary dynamics of the cryptocurrency market,” Royal Society Open Science, vol. 4, no. 11, November, 170623, 9 pages, 2017

[2] G. Hileman and M. Rauchs, “Global Cryptocurrency Benchmarking Study,” Cambridge Centre for Alternative Finance, 2017.

[3] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, A peer-to-peer electronic cash system, Bitcoin, 2008.

[4] J. Barrdear and M. Kumhof, “The Macroeconomics of Central Bank Issued Digital Currencies,” SSRN Electronic Journal.

[5] S. M. Raju and Ali Mohammad Tarif, “Real-Time Prediction of bitcoinprice using machine leraning techniques and public sentiment analysis”, Computer Science, International Islamic University, Malaysia, 2020.

[6] D. Sha and K. Zhang, “Bayesian regression and Bitcoin”, Annual Conference on Communication, Control and Computing, 2015.

[7] Ruchi Mittal, Shefali Arora, M P S Bhatia, “Automated Cryptocurrencies price prediction using machine learning”, ICTACT Journal on Soft Computing, vol. 8,2018.

[8] Sean Mc Nally, Jason Roche, Simen Caton, “Predicting the price bitcoin using machine learning”, IEEE conference, 2018.