**By:**

**Radhika Garg 401/2020**

**Harshit Jain 402/2020**

**Machine Learning Models for Predicting and Comparing Cryptocurrency Prices and Returns**

harshit jain

2021

**ABSTRACT**

Cryptocurrency has emerged as the latest financial assets. Its ability to facilitate digital barter has been met with both support and scepticism. This paper is an attempt to explore Machine learning abilities for the purpose of establishing models which can help identify price patterns and return potentials of cryptocurrencies in near future.

**INTRODUCTION**

Cryptographic currency is a milestone in the digital transformation of economies. It is not just an advanced peer-to-peer transaction system but an avenue of a promising new technology, the blockchain.

Invented by pseudonymous developer or group of developers ***Satoshi Nakamoto***, Bitcoin was the first cryptocurrency to have come into existence in the year 2009 and since then a myriad range of them have evolved in digital space. They are virtual assets which are secured using cryptography making it nearly impossible to counterfeit or double spend them.

***Cryptography*** is a collection of encryption algorithms which allow exchange of information only among intended users via codes. Such currencies use crypto encoding for the purpose of generating the public private ledger keys and also for the purpose of mining.

***Blockchain*** is a special type of electronic database where information is initially stored as blocks and later on chained together in chronological order where each entry is time stamped. Thus, the Bitcoin blockchain is just a specific type of database that stores every Bitcoin transaction ever made and collectively stores the data over a large network of peer to peer systems across world making it ***‘decentralized’***. Each node of the chain stores the data of the previously nodes so that in case there occurs an error it gets to cross reference and correct itself looking up the same stored information with other nodes. Thus, no single node holds the power to alter information in the network making transactions ***‘irreversible’.***

Cryptocurrencies have seen drastically increased capitalization due to lack of confidence in traditional financing systems and piqued the interest of governments, investors, and academia. But at the same time, they are extremely difficult to gauge in comparison to the fiat currency. They incorporate qualities of both a standard financial asset and a lot of speculative value. This in turn makes trading them a tedious task. Ergo, this research is an attempt to explore Machine learning abilities for the purpose of establishing models which can help identify price patterns and return potentials of cryptocurrencies in near future.

**LITERATURE REVIEW**

In the paper “Predicting Crypto Currency Prices Using Machine Learning and Deep Learning Techniques “ (Vaddi, Neelisetty, Vallabhaneni, & Prakash, 2020) several Machine Learning models were explored to provide an informative analysis of future market prices of cryptocurrencies using historical data. The models used were (Linear regression, Recurrent Neural Network(RNN), Long Short Term Memory (LSTM)).The tested hypothesis was whether cryptocurrency’s price can be defined as a function of other related attributes or not. The researchers commenced with Linear regression, a model where it was explored if prices were a function of independent vectors namely, blockchain size, number of wallets, hashing rate, etc for the last 365 days. The mean squared error (MSE) was optimized using Gradient Descent. Next RNN strategy was implemented as it is capable of storing the state of features from previous steps and can give enhanced predictions. But it was upgraded to an LSTM because RNN suffers a loss in performance when it has to memorize a lot of information. LSTM uses “cell state” to propagate information and solves this issue. Feature selection was done using Reduced Feature elimination (RFE). Also, nodes of layers were specified low (3 in the case) to prevent overfitting whereas training was done for 100 epochs to obtain 96.2% accuracy. Thus, it was concluded that LSTM was the best strategy.

In their study ‘A Research on Bitcoin Price Prediction Using Machine Learning Algorithms’ (Reddy & Sriramy, 2020) performed data scraping of bitcoin prices and trends, and predicted crypto currency prices using machine learning algorithms to select the best suitable model. They tested models like K-nearest Neighbors (KNN), Ridge regression, polynomial regression, linear regression and Random forest method to predict prices or future trends. Finally, all model’s accuracy was recorded and it was noticed that Linear regression model was the best fit. And using that, LASSO could be implemented to predict the time complexity reduction in bitcoin price. The machine learning algorithms can improve the future idea of crypto currencies. That will improve the market price of global investments.

(Phaladisailoed & Numnonda, 2018) in their study ‘Machine Learning Models Comparison for Bitcoin Price Prediction’ attempted to predict future price by using 1-minute interval trading data on the Bitcoin exchange website Bitstamp, prices of bitcoin were collected for past 6 years and used for prediction of future value of Bitcoin. The features used for prediction are open; close; high; low; weighted price; volume (BTC); volume (currency) and timestamp. These features are used to apply machine learning tools like Theil-Sen regression, RNN, Huber regression, Long short-term memory (LSTM) & Gated Recurrent Unit (GRU). And the conclusion is GRU shows the best accuracy of 99.2% R-squared value and 0.00002 MSE. In addition, the selected features may not be enough to predict the Bitcoin prices since various factors, such as the reactions from social media, policies, and laws that each country announces to deal with digital currency can all contribute to the rise and fall of Bitcoin prices.

(Raju & Tarif, 2020) in their study ‘Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis’ and tried to predict prices of crypto-currencies through sentiment analysis. The aim of their study was to compare between ARIMA and LSTM models to find which is the most efficient algorithm for predicting bitcoin price to ensure less risk and more profit for investors. The tools used were Tweepy to facilitate tweet collection from Twitter’s API; Python for implementing machine learning tools and Coinmarketcap, Bitstamp, Coinbase, and Blockchain Info for historical data collection of prices. This study used the daily price fluctuations of the Bitcoin to investigate the model's predictability with hourly price fluctuations in the future. And to create a model that can be used to successfully predict the price including the external factors.

(Kumar, 2018) created ‘A regression model for crypto-currency price’ and discussed the various factors affecting value of crypto-currency like types of crypto-currencies, mining, economic and legal aspects to determine the important factors that affect the price of cryptocurrency. Then data is extracted from coinmatrics.io with timestamps and price of cryptocurrency which is used for data modelling. Four most important crypto currencies (BTC, LTC, ETH and XRP) historical data was used for analysis. Various models were tested to check the fitness of data and concluded that a quadratic regression model using transaction volume can be used collectively for predicting price of all the four crypto currencies.

**OBJECTIVES**

* To build Machine Learning models which are capable of predicting Cryptocurrency prices.
* Compare returns of the top performing crypto assets namely Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Ripple (XPR) and Cardano (ADA) to reduce risk and generate profits when used to manage portfolios using time series analysis.

**METHODOLOGY**

**About the data:**

Machine learning models in this research use cryptocurrency transaction data from the following sources:

1. *BitInfoCharts*: It is a real time data tracking website where various cryptocurrencies can be compared on parameters such as price, market capitalisation, transactions, block-count, hash-rate etc.
2. *Crypto Data Download*: It is a historical data repository where data can be fetched from various exchange platforms to trade cryptocurrencies such as Binance, Bitstamp, Zaif etc. The research here uses data from Binance.
3. *Investing.com*: It is a website with both access to historical and real time data besides market analysis of assets including all cryptocurrencies.
4. *Blockchair*: The website provides informative analysis of cryptocurrencies using interactive charting tools along with historical data.

**Factors analysed:**

1. *Date*: We have taken daily basis data from beginning of January 2018 to March 2021.
2. *Price(in*USD): This is the closing stock price of cryptocurrency at the end of the day.
3. *Open*: Opening price of the cryptocurrency.
4. *High*: highest price of the cryptocurrency on that day.
5. *Low*: lowest price of the cryptocurrency on that day.
6. *Volume*: Number of coins traded in 24-hour cycle.
7. *Change%*: Percentage change in price in 24 hour cycle from open to close.
8. *MarketCap(in USD*): The total capital value of the respective cryptocurrency.
9. *Tradecount*: No. of transactions made in 24-hour cycle.

The data under consideration spans over the period of *March 2018 to March 2021*. For price prediction the data has a daily frequency with 1161 data samples. For return analysis we have considered monthly average of *change%* (percentage change in price from open to close) with 37 data samples.

Data obtained has been collated in the form of *excel* files. Analysis has been conducted using *Python programming language*. For ease of use and a robust analysis the research tends to work upon some top cryptocurrencies namely *Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Ripple (XPR) and Cardano (ADA)*.

**Machine Learning Models explored:**

Regression variants (Linear, Polynomial and ElasticNet): Regression analysis is a predictive modelling technique that analyses the relation between the target or dependent variable and independent variable(s) in a dataset. We will test the following regression models to find an optimal fit.

1. *Linear* *Regression* model consists of a predictor variable and a dependent variable related linearly to each other. It is the simplest form of regression.
2. *Polynomial* *Regression* is a technique to fit an nth degree polynomial relation between independent and dependent variables.
3. *Elastic Net Regression* is a regularized regression variant which over combines both L1 and L2 formats( lasso and ridge) when one is dealing with highly correlated independent variables.
4. *K Nearest Neighbour (KNN*): It is an intuitive method to approximate ‘feature similarity’ association among independent variable by observing the behaviour of the data points in the same neighbourhood.

1. *Time Series Analysis*: A time series is a sequential set of data points, measured typically over successive times. Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Here we will consider using Facebook Prophet. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality. It is robust to missing data and shifts in the trend, and typically handles outliers well.

**ANALYSIS**

**1) For price prediction**

**EDA**(*Exploratory Data Analysis*):The past trends of prices of different cryptocurrencies have been visualized as follows:

A picture containing chart

Description automatically generated A picture containing chart

Description automatically generated

*Cardano* has shown a tremendous increase in price. Trend is that of sudden rise with a stable low value in past. Also, it saw sudden decline initially.

*Bitcoin* has shown a tremendous increase in price. Trend is that of consistent rise with a stable low value in past.

Chart

Description automatically generated

*Ethereum* has shown a tremendous increase in price. Trend is that of consistent rise with a stable low value in past.

Chart

Description automatically generated

*Litecoin* has shown increase in price. Trend is that of fluctuating values of price.

Chart

Description automatically generated

*Ripple* started with very high price value but has consistently seen declined prices.

Also, it is evident from the price scales that bitcoin is the most expensive currency followed by Ethereum, Litecoin, Cardano and Ripple.

**Feature Processing:** Now we want to analyse the correlation between the different factors through *heatmap*.



For all crypto assets under question according to the heatmap, price has high positive correlation with open, low and high. Despite the fact these 3 features will not be considered for further analysis as they are variants of Price only. Volume, change percent and return have low correlation and are thereby dropped. *Market Capitalisation* and *Trade Count* are selected due to their high correlation with price. However, one might suggest that there is an *autocorrelation* between MktCap and trade count but it can be ignored as we have already zeroed out only 2 independent influencers.

Normal distribution of dependent variable is an important assumption for building any statistical models. On conducting *normality test* for price it is found that *data is not normally distributed as seen by very small p-values*. Therefore, we perform *Box-Cox transformation* of these variables so that it closely resembles a normal distribution. It is given as (y^l-1)/l if l != 0 and log(y) if l =0, where l=lambda varying as -5 to 5. The lambda chosen is such that it minimizes standard deviation to form a normal distribution.

**Model Metrics**: The data was *split* into test and train set with a 30% and 70% ratio. We tested out many models to predict the future prices of various cryptocurrency. The main determinants of evaluation are root mean squared error (*RMSE*) and *R-squared*. RMSE is the standard deviation of the residuals that is a squared weight of deviation of prediction from actual values. R-squared is a goodness-of-fit measure which denotes how much of the variability was a model capable of explaining.

**Visualising and Evaluating the Model fits:**

The actual price values vs predicted values for each model helps us visualize our model fit as in the best one will create a linear moving line depicting that there was not much difference and thus low errors.

**Chart, scatter chart

Description automatically generated**

*Linear regression* poor fitted all the currencies despite it was able to gauge the overall broader trend.

**Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated**

Polynomial Regression was conducted after finding out the best degree via iteratively generating RMSE. It was found that for all currencies it was degree =2 which gave the lowest errors. So, after fitting the model at d=2 it was found that irregular observations and learning path were generated.

Chart, scatter chart

Description automatically generated

*ElasticNet regression* also poor fitted all the currencies but was able to gauge the overall broader trend.

Here the *l1, l2 regularization* tradeoff affect was measured on various values (*l1\_ratio= 1,0.5,0.3*), none of which made any improvement in the R2.

Chart, scatter chart

Description automatically generated

*KNN regressor’s* actual vs predicted indeed gave very evident closeness in values for all the cryptocurrencies. Its performance was measured for a consideration of *10 neighboring* data points.

**Tabular Summary of Model Statistics:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Currency*** | ***Metrics*** | ***Linear***  ***Regression*** | ***Polynomial Regression(d=2)*** | ***ElasticNet***  ***Regression*** | ***KNN*** |
| ***BTC*** | ***R2*** | *0.6460* | *0.6460* | *0.6447* | *0.9975* |
| ***RMSE*** | *0.0039* | *0.0039* | *0.0039* | *0.0003* |
| ***ADA*** | ***R2*** | *0.5230* | *0.5230* | *0.50207* | *0.9974* |
| ***RMSE*** | *1.7546* | *1.7546* | *1.8060* | *0.1456* |
| ***ETH*** | ***R2*** | *0.7004* | *0.9126* | *0.6773* | *0.9970* |
| ***RMSE*** | *0.0198* | *0.0109* | *0.0207* | *0.0021* |
| ***LTC*** | ***R2*** | *0.8554* | *0.0968* | *0.8462* | *0.9916* |
| ***RMSE*** | *0.0292* | *0.0134* | *0.0295* | *0.0074* |
| ***XRP*** | ***R2*** | *0.5653* | *0.8386* | *0.5667* | *0.9922* |
| ***RMSE*** | *0.7396* | *0.4489* | *0.7209* | *0.1047* |

**Observations:**

Both are graphical and statistical findings indicate that *KNN* has not only provided a great fit model but also outperformed other algorithm fits with an R-squared standing at *99% and very low RMSE.*

**2)For Return prediction**

**EDA**

Chart

Description automatically generated

So returns seem to move randomly and make both profit and loss. But intricate analysis leads to following insights:

* *Cardano seems to be the most profitable.*
* *Ethereum and Litecoin follow suit as second best options.*
* *Ripple is fluctuating highly within profit loss ranges and seems risky.*

**Model Metrics**:

We have considered the *prophet* interface in python and *plotly visualisations* to create interactive graphs as seen below. Also, a future forecast data-frame for *next 24 months* has been considered for prediction. Later the model fit has been evaluated via *R-SQUARED* and *MSE* statistics.

**Visualising and Evaluating the Model fits:**

* **Bitcoin**

Chart

Description automatically generated

* **Cardano**

Chart, line chart

Description automatically generated

* **Ethereum**

Chart, line chart

Description automatically generated

* **Litecoin**

Chart

Description automatically generated

* **Ripple**

Chart, line chart

Description automatically generated

**Tabular Summary:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Currency*** | ***Observed Trend***  ***2021-2023*** | | ***Predicted Return as in Dec 22*** | ***R2*** | ***MSE*** | ***Rank*** |
| ***BTC*** | 0 – 0.03 | Fluctuating within positive ranges | 0.0319 | 0.66 | 1.75e-5 | 2 |
| ***ADA*** | 0 – 0.05 | Fluctuating within positive ranges | 0.0439 | 0.71 | 6.26e-5 | 1 |
| ***ETH*** | -0.01 – 0.03 | Positive to negative fluctuation | 0.0094 | 0.70 | 3.20e-5 | 3 |
| ***LTC*** | -0.015 – 0.035 | Positive to negative fluctuation most of them negative | 0.0432 | 0.63 | 3.05e-5 | 4 |
| ***XRP*** | -0.03 – 0.04 | Most negative fluctuation | -0.0062 | 0.67 | 5.12e-5 | 5 |

Time series analysis *R2 score is averaging at 0.68* which can be considered fairly well owing to the fact that the returns have indeed fluctuated highly. Considering investing options for the crypto assets .*Cardano* emerges as the best bet with highest returns followed by *Bitcoin*. *Ethereum and Litecoin* are predicted to have some potential for loss so invest with pinch of salt. Investment in *Ripple is highly risky* as it can even make losses of great amount.

**RESULTS**

*KNN regressor* was found to be the best model to predict prices based on internal factors such as Market capitalisation and Trade count with a very high accuracy as seen by the R2 score of 0.99.

*Cardano and Bitcoin* stands out as the most sought after investment option whereas investing in Ripple is equivalent to a gamble.

**CONCLUSION**

This research was aimed at exploring machine learning capabilities in the space of cryptocurrency price and return prediction. The data was of recent origin. But Crypto assets are known to be simultaneously affected by myriad of external factors such social media sentiments, policies and regulations, and country laws which are announced to deal with digital currency. All this can contribute to the rise and fall of prices and variations in returns. Therefore, it is suggested that as further research work *sentiment analysis* can be done to gauge the affect of external factors.

# References

Kumar, S. A. (2018). A regression model for crypto-currency price. *International Journal of Statistics and Applied Mathematics 2018; 3(3)*, 120-125.

Phaladisailoed, T., & Numnonda, T. (2018). Machine Learning Models Comparison for Bitcoin Price Prediction. *10th International Conference on Information Technology and Electrical Engineering.* ICITEE.

Raju, S., & Tarif, A. M. (2020). Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis.

Reddy, L. S., & Sriramy, D. P. (2020). A Research On Bitcoin Price Prediction using A Research On Bitcoin Price Prediction . *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 04*.

Vaddi, L., Neelisetty, V., Vallabhaneni, B. C., & Prakash, K. B. (2020). Predicting Crypto Currency Prices Using Machine Learning and Deep Learning Techniques. *International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, No.4*.

Wicklin, R. (2019). *The Theil-Sen robust estimator for simple linear regression*. Retrieved from blogs.sas.com/: https://blogs.sas.com/content/iml/2019/05/28/theil-sen-robust-regression.html