CRYPTOCURRENCY TIME SERIES FORECASTING USING MACHINE LEARNING AND DEEP LEARNING

AVLEEN SINGH
M.SC DATA SCIENCE

Final Thesis Report

JUNE 2022

DEDICATION

I dedicate my thesis work to my fellow professors, my thesis supervisor, and my mentor for their countless hours and push to complete my project. Their words of encouragement have never left my side.

I also dedicate this work and special thanks to Mrs. Charanjeet Kaur for her special attention and valuable feedbacks. She helped me develop my writing skills and my gratitude for those who spent countless hours in proofreading and supporting me to excel my leader dots.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Mrs. Aayushi Verma for her guidance, constant support and encouragement along with her several suggestions and comments that helped me complete this project and improve the presentation and writing of this study.

I wish to thank all my fellow professors who were more than supportive with their valuable time and expertise. A special thanks to my mentor Mrs. Aayushi Verma for her countless hours of reading, encouraging, and reflecting, and most of my patience in the whole process. Thank you all for agreeing to help me in this thesis.

I would like to thank and appreciate my professors for allowing me to do this research and providing any help needed. Special thanks goes to staff members for their continuous effort.

Finally, I would like to thank the administrators in the LJMU university who helped me with this thesis. Their willingness and special attention to give feedback helped me complete this project.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADF – Augmented Dickey-Fuller

ANN – Artificial Neural Network

AR – Auto Regressive

ARIMA - Auto Regressive Integrated Moving Average

ARMA – Autoregressive Moving Average

BART – Binary Auto Regressive Tree

BCH – Bitcoin Cash

BTC – Bitcoin

BTG – Bitcoin Gold

C&RT – Classification & Regression Trees

CBOE – Chicago Board Options Exchange

CvaR – Conditional Value at Risk

DL – Deep Learning

ETH – Ether

GAM – Generalized Additive Model

KNN – K-Nearest Neighbor

KPSS – Kwiatkowski-Phillips-Schmidt-Shin

LR - Linear Regression

LSTM - Long Short-Term Memory

MA – Moving Average

MAE - Mean Absolute Error

MAPE – Mean Absolute Percentage Error

MCDM - Multiple Criteria Decision-Making

ML – Machine Learning

MYANPIX - Myanmar Stock Price Index

NSE – National Stock Exchange

OHLC – Open High Low Close

RF - Random Forest

RMSE - Root Mean Square Error

RNN - Recurrent Neural Network

SVM – Super Vector Machine

SVM - Support Vector Machine

USD – United States Dollar

VaR – Value at Risk

Abstract

Bitcoin is one of the most anticipated and popular cryptocurrencies. This type of cryptocurrency is extremely hard to predict and track. It also has no correlation with market conditions. So, predicting its price is a tedious affair. Bitcoin is the first currency in the cryptocurrency market which has achieved great traction. Time-series analysis is the best way to predict the stability and market price of cryptocurrency based on Machine Learning. The future ups and downs of Bitcoin can be predicted with time-series analysis. For time series analysis, we have used FB Prophet, ARIMA, XG Boost and LSTM techniques. We have evaluated these machine learning models on parameters like Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and R-squared score. Bitcoin is a highly volatile asset. So, making and generalizing long-term predictions based on few ML models is not so accurate. However, we have used multiple models to make this research as helpful as possible for the investors and help them be able to choose the best model and predict the prices of Bitcoin in advance.

Keywords – "*cryptocurrency, time series forecasting, LSTM, Facebook Prophet, ARIMA, XG Boost, Root Mean Square Error, Mean Absolute Error, R-squared score, Bitcoin"*

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Background

With respect to different predictions and volatility of prices of cryptocurrency, there are several external and internal factors that can influence cryptocurrency values "(Derbentsev et al, 2019; Mittal et al, 2019; Mallqui & Fernandes, 2019; Parizi et al, 2018)". There are three external factors that affect crypto prices, i.e., market (trend, popularity, speculations, etc.); political (ban, legalization, etc.); and macro-financial (exchange rates, equity, interest rate, gold price, and other policies). Some other major factors are cost of doing transactions, supply and demand hash rate, compensation scheme, changes in circulation (Gibert Llauradó et al, 2020; Rathan et al, 2019). Bitcoin, Litecoin, Monero, Ethereum and Dash are some of the major cryptocurrencies whose values are affected by variables related to the crypto market, such as trade volume, beta volume, short- and long-term uncertainty, and volatility (Sun et al, 2020).

Cryptocurrency depends on the digital bookkeeping Blockchain model. Blockchain protects the privacy and user data with access management systems (Gurdgiev & O'Loughlin, 2020; Fang, Su & Yin, 2020). It is a connected and decentralized structure known for its native "data alteration resistance" but poor data formatting is highly responsible for lack of search query (Singh et al, 2019; Mehtab & Sen, 2020; Sattar et al, 2019). Abraham (2020) suggested an open-source framework for blockchain integration, ChainSQL, with database, in which, Blockchain is distributed, decentralized, and audible (Mohapatra, Ahmed & Alencar, 2019; Lahmiri & Bekiros, 2019). Blockchain relies on P2P, a decentralized network which mines currencies and handles transactions autonomously (Livieris et al, 2020; Chowdhury et al, 2020). All of the transactions of bitcoin are recorded in blocks in an open directory which is known as "blockchain" (Chivukula & Lakshmi, 2020; Anghel, 2021).

This authentication is performed in a trustless environment without any central authority to transfer funds between the recipient and the sender (Chen & Hafner, 2019; Yiying & Yeze, 2019; Poongodi et al, 2020). There is a time-series issue in advancement of cryptocurrency and trading (Kwon et al, 2019). It varies from traditional markets due to its high volatility and digital nature (Borges & Neves, 2020). There are also several causes behind changes in cryptocurrency price patterns and they are not easy to understand (Koker & Koutmos, 2020; Mikhaylov, 2020). But it is important to have an efficient mechanism to predict cryptocurrency prices for authorities and investors.

Technology has really transformed our lifestyle, work, and transactions. Financial institutions and economies are going digital across the world at an unexpected rate (Mudassir et al, 2020). The Fintech revolution has put the traditional financial system to shame. The emergence of cryptocurrencies is just another development in this domain and Bitcoin is the most anticipated digital currency which has a capitalization of a whopping $600 bn, which would just grow significantly in coming years. This cryptocurrency has become popular because of its great volatility and decentralized nature (Nakamoto, 2008; Baur & Dimpfl, 2017).

It delivers significantly higher returns to the investors than that of traditional investment or banking schemes. In addition, the price is measured fully by the public and the government is not involved in it. The market doesn't control the prices of cryptocurrencies. So, it is very important to conduct technical analysis to determine the price expectancy of bitcoin over a certain period. External economic data is not feasible for technical analysis because only past performance is used for prediction of Bitcoin prices (Abu-Mostafa & Atiya, 1996).

Researchers have recently used ARIMA and LSTM frameworks for forecasting Indian stocks' prices for 5 months and found better results with LSTM in comparison to ARIMA (Selvin et al, 2017). Li et al (2019) achieved 50% accuracy with "Gradient Boosted Tree Model" to capture public sentiments and data from Twitter for 3.5 weeks and each tweet was analyzed. According to a study conducted by Wint, FB Prophet had less error rate than ARIMA for daily, weekly, and monthly forecasts. The closing price of "Myanmar Stock Price Index (MYANPIX)" was predicted with both frameworks (Chan, 2020).

1.2 Problem Statement

Research on forecasting cryptocurrency prices has just begun. There are various studies related to various perspectives and methods to analyze crypto transactions. Some studies have investigated factors influencing the adoption of cryptocurrency while some authors tested ML techniques to predict cryptocurrency prices. Some have come up with various models to analyze cryptocurrency trends (Sun et al, 2019). ANN has been observed to be one of the widely used prediction tools in the finance domain. It can perform different tasks without compromising accuracy, such as prediction, classification, or pattern recognition.

ANN is a robust universal approximator for a continuous function. It is possible to approximate such functions with accuracy. The crypto market has been through a lot of ups and downs over the years. It is vital to know the trend because different coins are being exchanged these days. This study has compared XG Boost, LSTM, ARIMA, and Facebook Prophet models and found the ideal model for forecasting and prediction of Cryptocurrency prices.

1.3 Aim and Objectives

This study is aimed to find out an ideal machine learning model that can predict or forecast cryptocurrency prices with utmost accuracy and support a favorable environment for all crypto stakeholders with support of other economic ecosystems. On the basis of this aim, the research objectives have been formulated as follows –

- To develop and form a favorable cryptocurrency ecosystem of trustworthy members
- To form helpful guidelines for policymakers and regulating institutions to achieve competitive and effective regulations and promote the fair use, development and integration of blockchain
- To recommend the best ways to develop, recognize, and promote mutual understanding and cooperation between crypto ecosystem participants and international or government organizations
- To develop time series-based forecasting method which analyzes time-series data to gather sensible results and statistics from the data

1.4 Research Questions

- How efficiently machine learning models (LSTM, ARIMA, XG Boost, and FB Prophet) can predict cryptocurrency prices?

1.5 Scope of the Study

This study would be helpful for investors to understand price fluctuations of bitcoin, whether it's the right time to invest, ideal cryptocurrency to invest, etc. It is also helpful to compare different frameworks and choose the best one to predict cryptocurrency prices with utmost precision and accuracy.

1.6 Significance of the Study

Researchers have observed experimental Time Series data at various intervals, such as daily, hourly, or every minute. Daily data of airline ticket sales is one of the best examples of time series data. Just because there is a time element involved in a series of events and it doesn't make a time series automatically, for example, date of significant disasters in airlines, which are spaced randomly, these shouldn't be considered in time series. These random processes are called "point" processes. This study would consider various important elements of time series like seasonality, noise, and trend.

1.7 Structure of the Study

Future predictions can be made on the basis of forecasting, considering present and past data (Kavitha et al, 2020). In this study, we are going to use a big dataset. Hence, Kernel will be split into Data Exploration and Time Series Analysis. The Time Series Forecasting will be done further with –

- "Long Short-Term Memory (LSTM)"
- XGBoost
- "Autoregressive Integrated Moving Average (ARIMA)"
- Facebook Prophet

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

This section is based on recent developments on predictions for "cryptocurrency time series forecasting" with comparison of four models to find out the best possible model. In this section, we will explore all the earlier studies done to predict cryptocurrency rates with different Machine Learning models and conclude the studies.

2.1.1. Cryptocurrency prices time-series forecasting with "Binary Auto Regressive Tree (BART)"

Derbentsev et al. (2019) used a machine learning model for short-term cryptocurrency price forecasting. They adopted a modified "Binary Auto Regressive Tree (BART)" model from standard regression trees and time series data. BART consists of "classification and regression trees (C&RT)" and ARIMA auto-regressive models. They forecasted from 5 to 30 days in the short-term using BART model for Bitcoin, Ripple, and Ethereum cryptocurrencies which are most capitalized in the market. They found their approach more accurate to forecast crypto time series using "ARIMA-ARFIMA" models both in "transition dynamics" and "slow rising" periods.

2.1.2. Using ARIMA, LSTM, XG Boost Classifier, and Facebook Prophet to predict "BTC Time Series" data

One of the most anticipated and popular cryptocurrencies, Bitcoin is a kind of digital asset which is not easy to predict and track. Along with it, there is no correlation between BTC price and market conditions. Hence, it is not easy to predict its locus and price action. Wadalkar et al. (2021) compared four models to predict "BTC Time Series" data. They achieved accurate forecasting through the models. BTC prices were significantly high from December 16 to 31, 2020 because of heightened demand. The prices have fluctuated to US$8000 in quantitative aspects. Despite these significant changes in prices, they achieved a model to attain "Root Mean Square Error (RMSE)" of US$43231.80 and "Mean Absolute Error (MAE)" of US$153.55.

Earlier studies used one or two approaches for predicting BTC price. Bitcoin is a very volatile currency. Hence, generalizing them or making predictions for the long term on a few models doesn't give accurate results. These researchers bridged this gap by working with several models and fragmented time jumps in small parts. They predicted for two days only and achieved outputs with minimal error rates. ARIMA was found to be the best model to predict BTC time series trends in future. It also considers various decompositions like "Sessional Trend", "Regular Trend", and "Residual Trend" for best results.

- **LSTM** - "Long Short-Term Memory (LSTM)" is a kind of deep learning model based on "Recurrent Neural Network (RNN)". It is used in different fields like speech recognition and "machine translation" and it is also designed to classify, forecast, and process time-series data (Fang et al, 2019). The unnecessary information is filtered in cell states and removed to build the LSTM model. Post that, data which should be processed is stored followed by a "Sigmoid layer" which is also a filter to take information in the input layer. On the basis of these inputs through the sigmoid layer, the final output is obtained from a range of tanh features.
- **ARIMA** - The "Auto Regressive Integrated Moving Average (ARIMA)" model is one of the most popular "Autoregressive Moving Average (ARMA)" variants. It is designed for time series forecasting, processing, and data analysis. Each of these models are denoted by "p", "d" and "q" characters, where the letter "p" suggests numbers of lagging in the model, "d" refers to the times of observing differences in raw observations" and "q" refers to the size of "moving average (MA)" window. ARIMA denotes the same expression for Seasonal data in non-seasonal information with "P", "D", "Q" and "m" parameters. Here, "m" refers to the number of periods of each season, "D" refers to a differencing term, "P" denotes "autoregressive aspect" in the model, and "Q" refers to the "Moving Average" aspect (Zhang, 2003; Wang et al, 2013).
- **XG Boost Classifier** - It is a "gradient boosting" algorithm which uses the Ensemble ML algorithm based on a decision tree. It is a very powerful, portable, and adaptable algorithm. It is ideal for tabular/structured data of small to medium sizes and also problems related to decision trees as it provides "parallel tree boosting". XG Boost uses a "gradient descent" algorithm with new models to minimize the loss. Hence, it is called "Gradient boosting". It is meant to reduce the L1 and L2 objective functions. It is made possible with the integration of

penalty and uncertainty. The convex loss function works on discrepancy between goal and expected results. New trees are brought after each stage of the training process to project the errors/residuals of earlier trees. These are combined to make a final prediction.

- **Facebook Prophet -** This model forecasts time series information generated by Facebook. The "Generalized Additive Model (GAM)" arranges non-linear patterns with weekly, annual, and usual seasonality. It is very effective to predict seasonal data for any pattern, whether it is missing or highly seasonal. A 3-part time series model is used by the Prophet with seasonality, trend, and holidays.

2.1.3. ML Models for Predicting Price Rate of Bitcoin

The first ever decentralized cryptocurrency, Bitcoin has gained immense popularity in the market. It is also very difficult to predict because its prices fluctuate constantly. Kumar & Sunny (2021) have applied a time-series analysis on the basis of Machine Learning to predict the stability and market value of Bitcoin. They used Facebook Prophet, ARIMA, and XG Boost and tested these techniques on "R-Square (R^2)" and "Mean Absolute Error (MAE)" parameters. The most accurate model was Facebook Prophet in this analysis with R2 score of 0.99 and MAE of 3.2 percent to predict BTC price rate.

BTC prices are non-stationary and show variations in statistics over time. This way, Aggarwal et al (2021) used "technical indicators" that cause fluctuations in price of Bitcoin as an input method to ML algorithms to forecast trends, so that investors can make informed decisions. They presented robust ML models, such as Facebook Prophet, ARIMA, LSTM and XG Boost to predict long-term price fluctuations of BTC. ARIMA and Facebook Prophet achieved a whopping 95.72% and 95.70% accuracy, respectively. These models performed better than others in the study.

2.1.4. Using AI for Analyzing Cryptocurrency Price

Cryptocurrency has been very important to reshape the financial system with its growing appeal and merchant acceptance. A lot of people are investing in Cryptocurrency. The uncertainty, dynamics, and predictability are widely unknown in Cryptocurrency which is risky for investments. It is important to know the features affecting the formation of value. Yiying & Yeze (2019) used cutting-edge AI frameworks of "Long Short-Term Memory (LSTM)", "Artificial

Neural Network (ANN)" and "Recurrent Neural Network (RNN)" to analyze dynamics of Ethereum, Bitcoin, and Ripple prices. They observed that ANN depends more on "long-term history" and LSTM depends more on "short-term dynamics' ' which proves that LSTM is more efficient to use vital information in previous memory and it is better than ANN. ANN is capable of achieving the same accuracy with enough historical data in comparison to LSTM. They provided a different insight that it is easy to predict crypto market price. The predictability might also vary as per the nature of the ML model.

2.1.5. Predicting Price of Ethereum with LR and SVM

Poongodi et al (2020) predicted the price of Ethereum, a blockchain cryptocurrency, with two ML models – "Support Vector Machine (SVM)" and "Linear Regression (LR)" using a time series, which includes daily closing prices of Ethereum. The filters are used with various weight coefficients for various window lengths in predicting crypto price. A cross-validation approach is useful in the training stage to build a top-performing model which doesn't rely on a dataset. They found SVM with 96.6% accuracy while LR with 85.46% accuracy in the proposed model. In addition, they added features to SVM to improve accuracy score up to 99% in the proposed model.

2.1.6. Using "Chaotic Neural Networks" of Deep Learning for Cryptocurrency Forecasting

Lahmiri & Bekiros (2019) implemented DL techniques for forecasting the price of Digital Cash, Ripple, and Bitcoin, digital currencies which have been traded the most. It is probably the first work which used deep learning to predict cryptocurrency prices. By experimenting with non-linearity, it is found that time series of all cryptocurrencies result in long memory, fractal dynamics, and self-similarity. The predictability is very high in the neural network of "long short-term memory (LSTM)" as compared to the general regression neural model.

The latter couldn't determine the hidden and non-linear patterns despite contamination and noise elements and they rely on "Gaussian kernels" ideal for only local approximation of "non-stationary" signals. Despite having higher computational burden in the LSTM model instead of brute force in recognizing non-linear patterns, deep learning was, at the end, observed as highly efficient to forecast chaotic elements of the market.

2.1.7. Predicting Stock Price with DL and ML Frameworks

Prediction of stock prices is a challenge to the researchers in the financial segment. Though it is not possible to accurately predict stock prices according to the "Efficient Market Hypothesis", there are studies which show that it is possible to forecast stock prices with reasonable accuracy by choosing the right variables and ideal predictive models to be developed with these variables. Sen & Chaudhuri (2018) presented an accurate and robust stock price prediction framework with ML, DL and statistical models.

They used daily stock price data at the interval of 5 minutes from India's NSE and aggregated such data to build the framework for stock prices. With a combination of deep learning and ML methods, this framework can model the stock price volatility precisely and it can be used to forecast stock price in the short term. There are 8 regression and 8 classical models and one model on DL approach which have been developed with data of Hero Moto and Tata Steel stocks listed in National Stock Exchange. They presented results on performance of such models.

2.1.8. Crypto Price Forecasting with Machine Learning Models

Over the years, cryptocurrencies and blockchain have gone through a significant change because of its significant capitalization and trading volumes in the market. Along with trading, these cryptocurrencies are being used and accepted for financial transactions these days. With the fluctuation in prices and rise in ROI, traders, investors and common public are more attracted for altcoins and bitcoin. Chaudhari (2020) implemented forecasting models which will get back with accurate crypto price predictions. They predicted Litecoin, Bitcoin, and Ethereum prices with conventional forecasting model for ARIMA timeseries, LSTM deep learning model and FB Prophet model. They found LSTM to perform better than ARIMA and Prophet model by analyzing three models.

The impact and occurrence of financial markets is ever rising on several domains like education, businesses, technology and jobs and also affecting several economic factors and sectors. Price movement of stock market and stocks has been one of the major causes for traders, investors, and scholars to develop a keen interest on this topic. It has also developed several models for price and behavioral prediction with various techniques like “Pattern Recognition”, “Statistical Analysis”, “Sentimental Analysis” and “Machine Learning” (Shah et al, 2019).

In financial markets, cryptocurrencies have literally revolutionized financial markets over the years and their profitability have brought them to the limelight. It is also possible to categorize the techniques for price prediction as it is much similar to stocks. Figure 2.1 refers to the taxonomy of most common prediction methods which have performed well already.

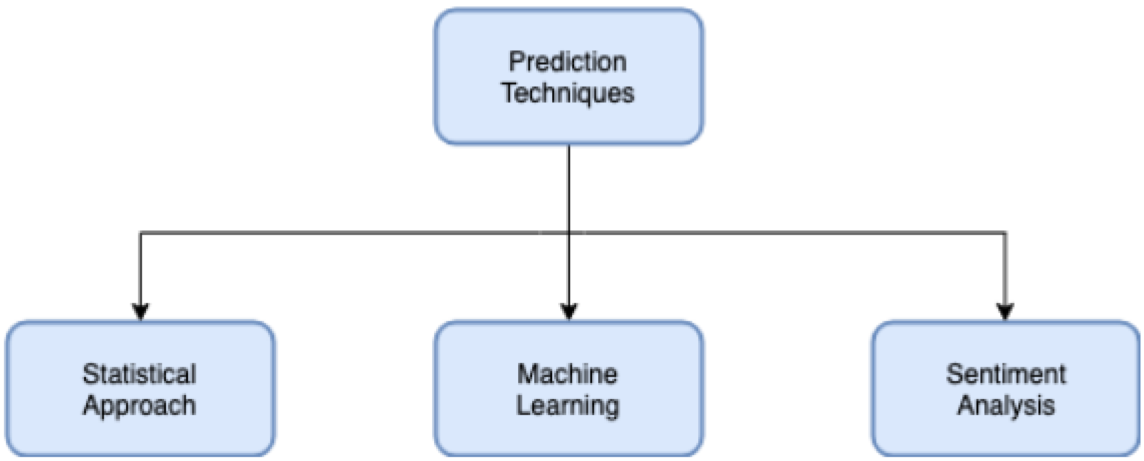


Figure 2.1. Popular Techniques for Crypto Price Prediction

Karakoyun and Cibikdiken (2018) and Saxena and Sukumar (2018) have employed LSTM and ARIMA statistical and deep learning models in their studies and achieved RMSE of 456.78 for LSTM and 700.69 for ARIMA models and RMSE of 93.27 for LSTM and 1146.07 for ARIMA models. However, LSTM has not been mentioned in both studies that it has achieved error rate of only 1.4% and proved more accurate than ARIMA.

Peng et al (2018) and Yu et al (2019) have adopted “GARCH models”, while Bitcoin’s price volatility was analyzed in the initial research to determine the volume’s impact on volatility from the asymmetrical perspective apart from the growth analysis of interest with Wikipedia and Google Trends views. The “SVR ML approach” and “GARCH model” were combined in the next study. The latter study found better performance of “SVR-GARCH” against all “GARCH benchmarks” as values of “MAE” and “RMSE” were lower for developed model with ML approach.

The “ANN-GARCH” model was proposed by Kristjanpoller and Minutolo (2018) along with forecasting volatility preprocessing in prices of Bitcoin and it was compared with “EUR-USD” volatility. They obtained MAE and RMSE analysis for several horizons and applied MCS test. The hybrid models were not superior in comparison to GARCH models for 5% significance.

Wu et al (2018), Lahmiri and Bekiros (2019) and Alessandretti et al (2018) have used LSTM as base model. But only LSTM was used by the second group of researchers and found better performance in LSTM as compared to GRNN, while other two studies also included other models. The results of common LSTM model were compared with “LSTM-AR(2)” by Wu et al (2018) and both models forecasted the price near the real values. The “LSTM-AR(2)” model resulted in lower value of MAE, RMSE, and MAPE. A total of 1681 cryptocurrencies analyzed by Alessandretti et al (2018) and XG Boost trees performed well in the results for around 5 to 10 days but more accurate prices were obtained from “LSTM-RNN model” in the long run. RNN is also used by Pant et al (2018) for prediction of Bitcoin prices with historical data for the next day. A sentiment analyzer is combined by this study which can differentiate between negative and positive tweets and found 77.62% accuracy with RNN model.

Ensemble, SVM and ANN algorithms were used by Mallqui and Fernandes (2019) to predict Bitcoin price directions. The regression results were used to improve price direction prediction as inputs. Around 10% of accuracy is presented in the results as per the given attributes. They obtained 1% to 2% of MAPE values regarding the minimum and maximum closing prices for Bitcoin. Hence, this model has been more efficient than others and suggested to improve predictions with economic and technical indicators.

Singh and Agarwal (2018) have applied four traditional machine learning approaches to predict bitcoin prices in which all the topographies were used separately for “Polynomial and Linear Regression” and “Polynomial Regression”, “SVR” and “KNN” were implemented for hyper-tuning with “grid-search logic”. KNN performed better than other models and achieved 0.00021 MSE. SVM, KNN, and LASSO were also used by Liew et al (2019) apart from “SGD”, “ElasticNet”, “Decision Tree”, “Bayesian Regression”, “ADABoost”, “Decision Tree”, “MLP” and “GTB” for predicting 30-day returns from cryptocurrencies. Around 50% to 60% of accuracy can be generated by almost all algorithms but 61% of accuracy has been obtained by LASSO for this model. According to the authors, low volatility in huge cryptocurrencies could be predicted more accurately than novel crypto assets with high volatility.

Sentiment analysis has been used to predict crypto prices with collection of tweets over time by various researchers like “Wimalagunaratne and Poravi (2018), Rahman et al. (2018), Bibi et al. (2019), and Steinert and Herff (2018)”. Various classification and regression algorithms have been used by Rahman et al (2018) and 89.65% of accuracy was obtained by Naïve Bayes and proposed prediction could be helpful with sentimental analysis in changing crypto prices, while “Topic modeling” was used along with “opinion mining” by Bibi et al (2019). Users’ concerns have been investigated by authors along with their sentiments at identified locations where use of crypto is paramount.

Historical and trading information, dependency on altcoins and Bitcoin, and public perception were assessed by Wimalagunaratne and Poravi (2018) and applied sentimental analysis and machine learning to elevate price prediction efficiency. Meanwhile, Steinert & Herff (2018) predicted altcoin returns with “sentiment analysis” with the application of “linear regression” as per the collection of 45-day data. Considering the results achieved, it is possible to predict Twitter activity and short-term sentiment bubbles. According to the authors, prediction accuracy may be improved further by modern prediction methods like “Neural Networks”. Prophet and ARIMA models were used by Yenidoğan et al (2018) for predicting crypto prices with bitcoin data gathered from the year 2016 to 2018 and results found that Prophet outperformed ARIMA with R2 of 0.94, while ARIMA could possibly achieve 0.68 as R2.

2.2. Summary

Due to the recent emergence of this phenomenon, academic literature based on cryptocurrency is quite limited in comparison to more traditional topics like “stock market” and usually go through the growth and decline in hype of cryptocurrencies, which are closely associated with steep price changes. Diverse topics have been explored as authors addressed various issues from the impact of crypto market on ecology due to energy consumption to their behavior as compared to mature markets to attempts to forecast their volatility, prices, bubble bursts and formations, and cyclicity.

Given the lucrative implications, price forecasting is definitely one of major aspects in cryptocurrency research. There are different methods used from correlation between crypto assets and time series like valuable metals, stock indices, or foreign exchange, to autoregressive ones like VAR, ARIMA and several extensions. Autoregressive methods are quite common and are unsurprising as they are de facto for modeling of time series information and they only need little bit of contributions for approximation. “ARIMA” required only the “lagged values” of forecasted series. Various studies have been done for giving information on crypto asset space with indicators. This study is aimed to fill the gap between various machine learning algorithms and cryptocurrency price forecasting.

CHAPTER 3

METHODOLOGY

3.1. Introduction

The raw dataset itself consists of price action for predicting the price action both as the input/output of the models and input variable for trading volume. Price action is used to forecast the main variable while its main predictor is its previous values. Trading volume refers to each currency's amount which is transferred across the owners during the specific time period and is supposed to have highest prediction potential along with action of historical price itself. Though it is also possible to predict the trading volume, it results in least useful data and data that could be achieved off it, including the price action volatility with its utilization as a proxy and is modelled well with volatility instantly. Hence, the initial dataset consists of two time series for both volume and price examined.

Cryptocurrencies display almost unexpected volatility unlike other assets in time series forecasting and it is also tied to very uncertain pricing, significantly low liquidity, which causes huge price fluctuations due to the buying and selling of negligible amounts of crypto assets in comparison to total supply and vulnerability to bubble speculations. The volatility is commonly seen followed by visual inspection of price development of several cryptocurrencies. Along with several issues of uncertainty for the investors, it also poses forecasting and modelling issues, especially for approaches where stationarity is needed.

Despite having high amount of intermarket cohesion, assets prove idiosyncratic drops and surges in price sometimes, which may show useful data and be forecastable and it is also supposed to be lost if a composite index or crypto asset was used to represent markets with "one price action time series". The properties also consist of exclusive qualities given fully in their subsets, which may affect their rating with processes which are apparent, more or less. These assets are used to have market entities which are very diverse by nature, while also using some top positions in the market with availability of sufficient information (along with reasonable age of asset for historical data to exist in the right volume), and relevant data about the order book depths of the assets.

While including various cryptocurrencies with various attributes could cause the rise in possibility of looking for highly forecastable time series and enhance the applicability of the work, a lot of assets don't have proper liquidity and call for "book depth". Their "price action" may have spikes in prices which show that the "order book" of the asset was emptied out well, leaving just extremely high orders of "buy or sell makers". Dealing at the given closing price in those situations is not the viable solution and there is little value offered by the price information.

Though every possession in this study should have proper liquidness for the buying or selling of asset by the depositors in the given time period, it is viable to manage this situation by including periodic snapshots of order book in the data, even though it causes inflated asset and snapshots may not be easy to achieve for the assets which are required. The distribution issue has widely been technical for a lot of cryptocurrencies. The chances of either pressure from the state on certain currencies occurring from their uncertainty or higher public interest with user-friendly distribution or the recent legal and public considerations may have great implications for the adoption rates and longevity of assets and their prices.

This section consists of important processes like "Data Selection", "Data Preprocessing", "Data Transformation", "Data Balancing", and evaluation metrics to evaluate ML performance and techniques. It shows a dataset and proposed ML algorithms for time-series forecasting of cryptocurrency prices.

3.1.1. Data Selection

We have collected "Bitcoin Historical Data" from an open-source platform, Kaggle, which consists of eight features described in the following table –

Table 3.1. Features of Bitcoin Historical Data

Feature Name	**Description**	**Format**
Timestamp	Time of each entry of instance that has been gathered from crypto stock market	DD/MM/YYYY
Open	It refers to open price of cryptocurrency on daily basis as per the timestamp	In Million US dollars

Close	It refers to the final price of cryptocurrency on the day of collection of data	-do-
High	It refers to the highest price of cryptocurrency on the day of data collection	-do-
Low	It refers to the lowest price of cryptocurrency on the day of data collection	-do-
Volume (BTC)	It refers to turnover in BTC prices	-do-
Volume (Currency)	It refers to turnover in currency exchange rate	-do-
Weighted Price	Shared price average of all investors in bitcoin	-do-

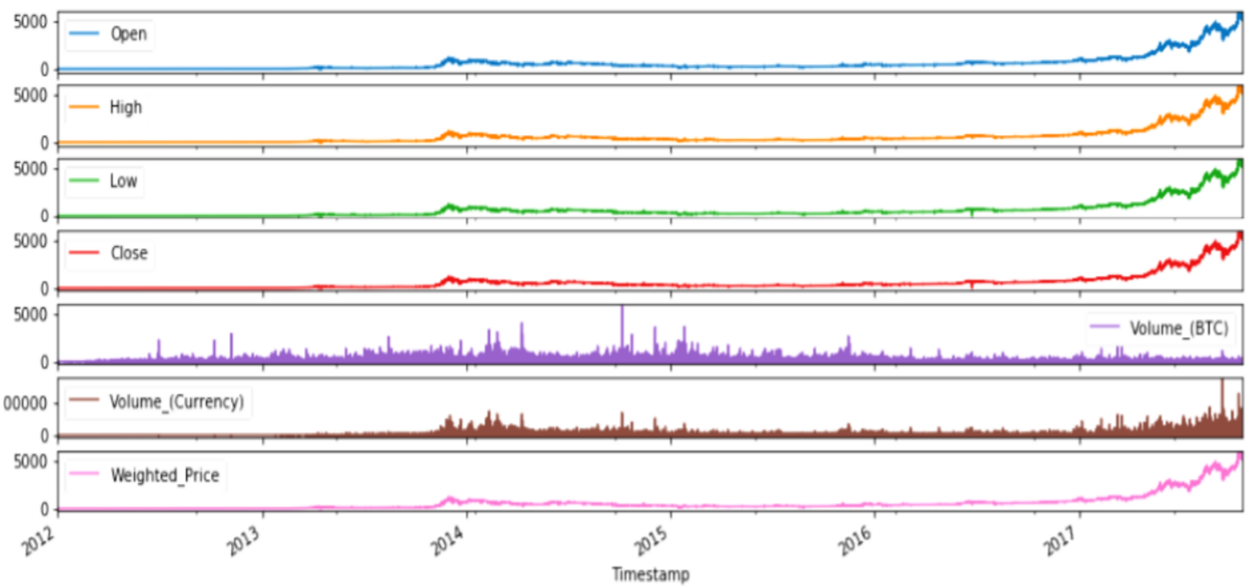


Figure 3.1. Exploratory Analysis of Historical Data

3.1.2. Data Preprocessing and Transformation

Preprocessing refers to filtering unwanted data off the dataset. This way, we have got rid of outliers, normalized features, and conducted important attributed analysis. Data transformation and pre-processing techniques for classification are “Outlier Treatment”, “Class Imbalance”, “Binning Continuous Variables”, “Skewness Correlation”, “Boolean Variable Mapping”, “Duplicates Removal”, “Data Normalization”, “Date Format Correction”, “Dummy Variable Creation”, “Test-Train Split”, and “Feature Scaling”. On the other side, data transformation and preprocessing techniques for time-series forecasting are “Differencing”, “Stationary Tests (ADF, KPSS, etc.)”, “Missing Value Treatment”, “Time Series Decomposition”, etc.

3.1.3. Data Modeling

The following machine learning algorithms are ideal for the proposed model to predict cryptocurrency price –

- **ARIMA** – The “Autoregressive Integrative Moving Average” model is used for determining and forecasting the future values on the basis of historical pricing. This model has historically been used for time-series forecasting where the combination of “Moving Average (MA)” and “AutoRegressive (AR)” models work to form “ARIMA” model (Karakoyun and Cibikdiken, 2018). The data becomes still for non-stationary time series in ARIMA with finite differencing. It usually consists of diagnosis, parameter estimation, and model identification.
- **LSTM** – The “Long Short-Term Memory Networks” are RNN with ability to explore long-term dependencies. It means they can easily remember the data for the long term in comparison to RNN, With the help of deep learning model, large architectures can be trained and become better against the errors (Saxena and Sukumar, 2018). Hochreiter and Schmidhuber (1997) have introduced the LSTM model with a network having several layers, including the output and input layer apart from either multiple or single hidden layer(s). It constantly learns several steps of time and “forget and remember gates” which help in determining the data to be passed on the basis of its strength and value (Karakoyun and Cibikdiken, 2018).
- **Prophet –** Developed by “Facebook’s Data Science Lab”, FBProphet model is used to forecast daily, annual, and weekly non-linear trends and it is based on additive model where such trends are fit. It uses a “modular regression” approach which enables selecting components with a forecasting issue and makes changes accordingly, while working with default parameters well. It also consists of measuring and forecast tracking system so that analysts can make changes and improve forecasts with “incremental adjustments”. Prophet is basically an adjustable and adaptable model to deliver scalable performance and analyze various time series (Taylor & Letham, 2018).
- **XG Boost –** This “gradient boosting” model uses ensemble ML algorithm which relies on decision tree. It is designed to be very adaptable, portable, and powerful. It is the ideal option for tabular/structured information of various sizes and also offers “parallel tree boosting” to

deal with decision tree-based issues. It reduces the loss with “gradient descent algorithm” while adding new models. Hence, it is called as “gradient boosting” algorithm. It is aimed to reduce “regularized objective function (L1 and L2)”. It can be achieved with the integration of uncertainty with the convex loss function and penalty term on the basis of discrepancy between the goal and expected outputs. New trees are brought after each training phase to project the errors/leftovers of earlier trees. These are combined to make final predictions.

This research will be conducted on the basis of the following methodology (Figure 2).

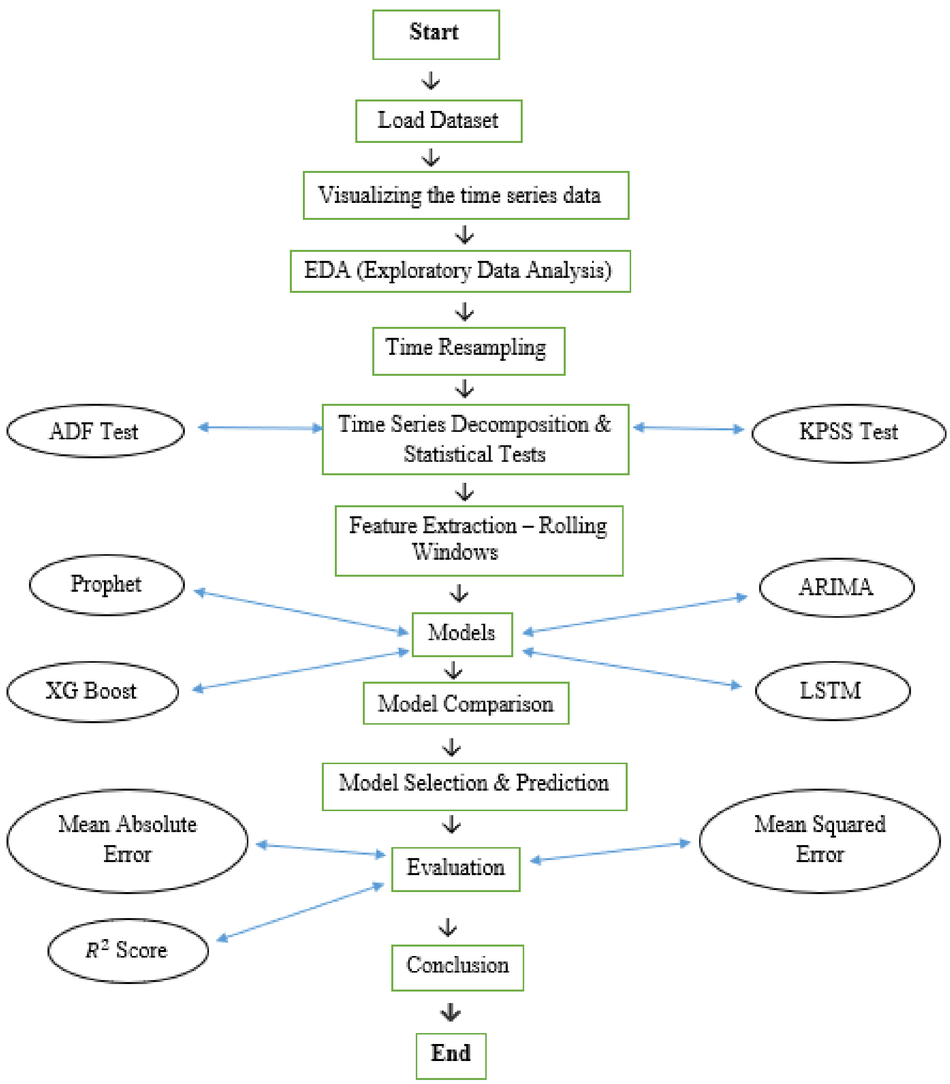


Figure 3.2. Flowchart of Proposed Methodology

3.1.4. Distribution Systems

The "proof-of-work" is probably and only the best distribution system where the provision of needed work is constantly known as mining and it still rules the market with different miners instead of the alternative method users. This is definitely associated to the circumstance that it is widely used still as a major "distribution approach" by leading crypto assets in marketplace capitalization, i.e. Bitcoin and Ethereum. The miner should provide computing power for proof-of-work to solve different cryptography tasks to choose the miner for verification of another block in the blockchain and achieve the reward. Then, the new crypto asset units are transferred to the miner's wallet. This distribution system is constantly scrutinized heavily for mining tools and power consumption. The GPUs constantly wok at their full potential and need a lot of power. In addition, economies of scale benefit the mining industry and prevent the real decentralization of crypto distribution.

Posing as the major "proof-of-work" alternative, "proof-of-stake" is more sustainable platform for delivery which helps asset owners to stake their holdings. The owner loses their control to use staked units for as long as they are staked for transactions, apart from transfers from one wallet to another of the same entity or person, but rewarded for this liquidity loss with an option to get more of the cryptocurrencies. The more of the asset is at stake, the higher the chance for the wallet to be selected for validation of the block and get the reward for the block. The minimum time needed for staking the asset varies as per the crypto assets. There is also no restriction on some asses on staking time. Some need several days or months. Ether is the most closely associated asset which uses "proof-of-work" mainly as its distribution mechanism, with the staking share supposed to grow soon. It is probably becoming the only system of distribution as it also covers "Binance Coin", which is exclusively distributed by staking.

Another recent addition to "distribution" is "proof-of-space" or "proof-of-capacity" but its market is not that mature. Its agreeable ecological outlook is often used as its "extenuating" add-on. It only needs specified virtual space to files that are deemed useless. Those files are existing and verified from time to time. If their existence is confirmed by the verification process, their holder may be rewarded for completion of block successfully as per the random process, where the chance of winning a reward is based on the dedicated space to the files by the miner associated with total allocated space for the network.

Though it seems to require less resources as compared to proof of work, it goes without saying that creating certain files may need special solid storage to lower its given lifespan significantly and it is also vulnerable to centralization. Some cryptocurrencies relinquish the distribution which relies on verification of block or other network services at the same time. Rather, they distribute only novel components across the marketplaces. These possessions are known as “pre-mined”. This system also has risk inherent from the fact that authors of the assets often have vast share of total assets to work equivalent to central bank or to sell the whole holdings simply for significant wealth, but might be destroying the asset’s reputation and value forever.

3.1.5. Cryptocurrencies

Bitcoin is one of the obvious cryptocurrencies to be modeled. Currently, it is the oldest, most publicized, and largest cryptocurrency in the market and has significant correlation with other cryptocurrencies. Its prototype also makes it simplistic in comparison to its recent counterparts which can be ironically expected to affect its modest strength in the extended run. Its value comes from the typical attributes of standard currency and it serves as a medium of exchange, unit of account, and a store of value. Though these functions are quite reduced in the meanwhile because of volatile nature of bitcoin and low distribution in common public, other cryptocurrencies also suffer due to such problems and result in added functionality. Bitcoin has been forked in majority of cases to deal with this rising threat of getting obsolete and created new cryptocurrencies like “Bitcoin Gold (BTG)” and “Bitcoin Cash (BCH)”. The BTC market consists of very reasonable amount of liquidity in comparison to other cryptocurrencies.

The historical price of bitcoin shows an inconsistent trend for a while where a bubble develops after significantly low activity. The price goes down to the levels which exceed the average levels of price of the earlier period of low activity rather than being disastrous with interpretation with latest peak price when a crash takes place after the bubble. The frequency of appearance of these bubbles is also increasing over time. Bitcoin price is quite stable in comparison to other cryptocurrencies, even though it is almost unexpected in volatility in traditional markets.

Figure 3.3. Bitcoin closing prices and “Exponential Moving Average (EMA)” for 50 days

Ether (ETH) is the next asset class associated with Ethereum, its overarching project. Bitcoin is supposed to be a fiat currency alternative and also consists of capabilities of a smart contract. This way, ETH holders can work on “decentralized contracts” and enable conditional ETH transfer to third party on the basis of existence of consensus or the given event. Suppose a person has planned to build a house. He has sent the advance to get the construction done, but he also does not want to get in trouble. This way, a smart contract is written to guarantee that the imbursement will be out only when the service provider builds the house and both parties agree to that. It is possible to use more complex contracts as cryptocurrencies like tokens are made with programming language. They use blockchain of Ethereum for added functionality.

In 2017, Ethereum has recorded huge returns with ca. $1350 from merely ca. $10, reaching all-time high. However, when crypto bubble bursts, the value was declined to merely ca. $83, which is still significantly higher than the price levels before bubble. From 2018 to 2020, the price also remained at distant levels from all-time high, followed by the growth in 2020 and 2021, which have led to a great spike.

Figure 3.4. Ether Closing Prices and EMA of 50 days

XRP is the next asset class named after Ripple, its parent company. Hence, it is also known as Ripple. It can resist scaling issues unlike the rest of cryptocurrencies mentioned above. Historically, the network of several other currencies couldn't have a lot of transactions taking place. Hence, asset transfers took a lot of time and severely affected medium of exchange for those currencies functionally. Since then, there is a decline in severity of this issue. But there are still opportunities to slow down these asset networks by huge volumes of transaction. XRP can highly resist this issue by design as the transactions of the network are verified by the list of validators approved by Ripple. Rather than a distributed ledger, the company operates some of them directly.

The validator nodes are also selected by the transaction parties they rely on despite the sanctioning by Ripple itself. Even though a lot of nodes are clogged, there are always chances to look for alternative validator. XRP is not completely decentralized in majority of crypto space as epitomized by Bitcoin. Along with the nodes which are partly centralized, most of its supply was mined earlier than the introduction of cryptocurrency to common public and a part of total amount is held by a limited number of people which has not been circulated yet. Most of them are the authors in the project. Hence, XRP is a cryptocurrency which challenges a lot of its core principles which are similar to most of the crypto spaces.

The price of XRP has witnessed a significant rise in the year 2017, when it grew from just around 0.5 cent to whopping $4 next year in January or over 60,000% per unit change in valuation. Even in the 2020 and 2021, this record has not been broken in positive environment. Eventually, XRP lost its position being the third cryptocurrency as per the market capitalization. However, it still remained the "non-negotiable asset" with various features. It is worth noting that the 2017 bubble is still not shown in Figure 3.4, as it was not the part of dataset because of availability of data.

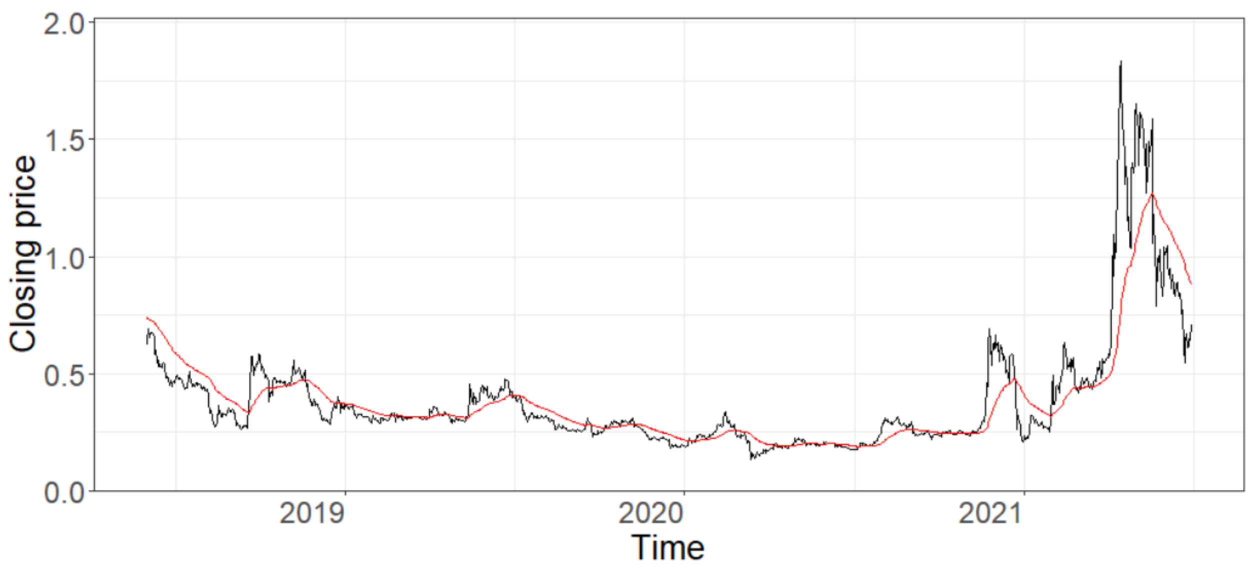


Figure 3.5. XRP Closing Prices with EMA for 50 days

"Binance Coin" is the last asset class here. The best part of this cryptocurrency is that it is not associated with its leading position or advancements in technology unlike other assets. Rather, it is tied closely with "Binance exchange" in China which was established in the year 2017, which is holding its topmost position among "cryptoasset exchanges" in trading volume. With the help of "Binance Coin" users can reduce their trading fee by around 0.1% to 0.05%, despite the traded asset until the BNB is held by the user and uses it for the same purpose on their own.

Hence, the exchange itself achieves a lot of its currency with such fees and has become a promising destination to remove a huge part of the components being held periodically, leading to a decline in entire source, and foremost to the development of the price theoretically in the horizon of the "long term". This way, "Binance Coin" is held as a money in the form of stocks of the company.

Public interest is also increased in Binance Exchange which must cause higher demand along with the burns for discount in trading fee. Initially, Binance Coin was a token which used to rely on Ethereum. Later on, it became an individual asset with a ledger.

The price of “Binance Coin” has seen a rapid growth by the end of 2017, apart from the rest of cryptocurrencies, even after a crash. Unlike several other cryptocurrencies, the price of BNB reached its skies drastically in the mid of 2019, when a lot of markets vacillated around the lower levels after crash and also reached significantly lower price than expected sometimes. Its ties to exchange might be the cause of this superior performance as a lot of cryptocurrencies were traded with Binance, which required an approach to reduce fees significantly. During the rest of bearish market, BNB has held its appreciable value and also witnessed significant growth in the year 2021.

The data for time series forecasting was taken from CryptoDataDownload, which sums various frequencies’ data for close and open prices and daily minimum and maximum value, apart from volumes. It provides individual exchanges’ prices, of which Binance is at #1 as per traded volume. There is also a chance for aggregated data to have virtually all or several exchanges. It is worth noting that this approach may consist of exchanges which don’t need “trading fees” in its group of data and it may subsequently have “trading bots” that sell and buy assets again and again, or even exchanges which fake the volume of trade right away. More and more traders are attracted to it which fakes liquidity as the volume is irrelevant and artificial at the end of the day. One can expect vital role of robotic arbitrage to keep prices under control across the exchanges for viable data on single exchange.

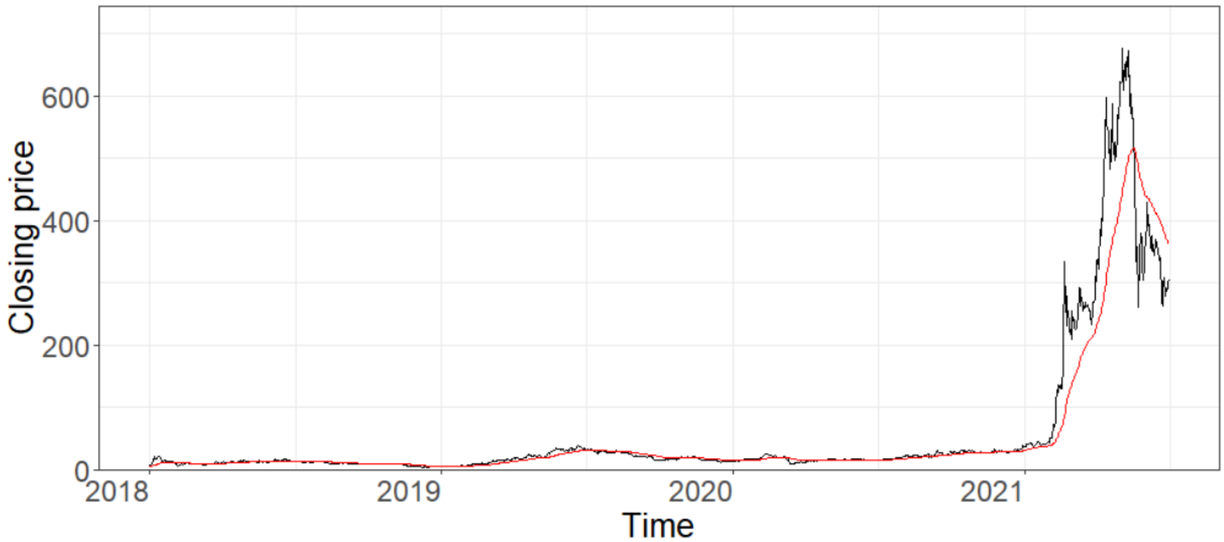


Figure 3.6. Binance Coin Closing Price and 50-day EMA

In further analysis, the choice of periods is random most of the time. The selection is done to ensure proper judgment of forecasting performance both in very highly volatile bubbles and when bubbles are calmer. The cyclical time series also leads to certain risks. The performance of the model might struggle in order to forecast the prices for the bubble as training set consists of "inter-bubble data" only. Each cycle is significantly different at the same time in its volume and price changes. While this problem is addressed by differencing addresses and logarithmic change, the models are used at the end for forecasting unexpected events.

Given the implications like huge changes in volatility and its cyclicality along with data availability, the period for testing and training machine learning models selected were from September 1, 2017 to June 30, 2021 for Ether and Bitcoin, having 3 quarters a year, including the fresh bubble for testing along with the last one since 2018. There is same terminal date for XRP and BNB data but they start on June 1, 2018 and January 1, 2018, respectively. For price action, closing prices have been used as proxy as they are supposed to be more robust in comparison to max and minimum price of the day, and similar to the opening price on the next period.

There are also average daily prices but they are not easy to accumulate for all currencies given above. A lot of data sources report only high/low prices, opening, and closing. It is not possible to be used separately to generate the true average. Given the frequency of time series, the only timescale used was daily data. For this choice, the reason behind is that there are often exchange fee issues for shorter frequencies with the special interest of the work in trading applicability. There is also a rise in daily changes though, i.e. up to 10% points as compared to the last period, to make fees look minimal.

The fees may decline in the market in the short term or even surpass the revenue in trading. Lengthier time is often suboptimal for both autotelic modeling and trading. The main cause behind is that it fundamentally limits the modelling data for training as given that older data than "circa 2015" is omitted as per the erratic behavior of previous markets. Around 60 periods have been left for monthly model training. In addition, the appearance of bubbles in crypto markets till now have lasted only for around 6 months. Longer time prediction could avoid serious points of interest completely from such periods, including the beginning and peak of the crash, which are virtually vital for both modelling and trading.

All the explored "time series" were changed "logarithmically" and changed for investigation once more. Unlike other approaches, ML models don't need stationarity and forecasting the price change is the main reason behind, especially during the period just after the last actual value. One can interpret the differenced and transformed dataset and represent price changes for the specified goal. The effect of huge price changes can also be mitigated with logarithmic transformation which take place semi-frequently in crypto markets.

3.2. Summary

This time series forecasting is mainly aimed to get useful results to help investors to make informed training decisions and these rely mostly on negative or positive price forecasts, instead of its size. The relative price changes are also important to deal with the major problem which is being discussed. The "time series" of values representing change for each series period is determined with respect to the period which is constantly preceding. It is applicable to both volumes and closing prices time series.

This study will compare the above models with one another and come up with the right technique for forecasting cryptocurrency prices. The “sklearn metrics” is known for its high accuracy as it uses various utility functions, losses, and scores to determine regression performance. The “Root Mean Square Error (RMSE)”, “Mean Absolute Error (MAE)” and R-squared score have been used to determine accuracy of all four models used in this study.

CHAPTER 4

DATA ANALYSIS

APIs of Poloniex (2019) and Quandl (2019) are used to extract data and data is processed with data cleaning, calculation, and aggregation of altcoin prices. Then, the prices of altcoin and bitcoin are merged into an individual data and data frame for "time-series forecasting". Then, the algorithms are put and models can be implemented. Then, one can evaluate the output as per the evaluation metrics for all models by comparing results to come up with best algorithm for time series forecasting. Figure 4.1 illustrates the whole process of implementation.

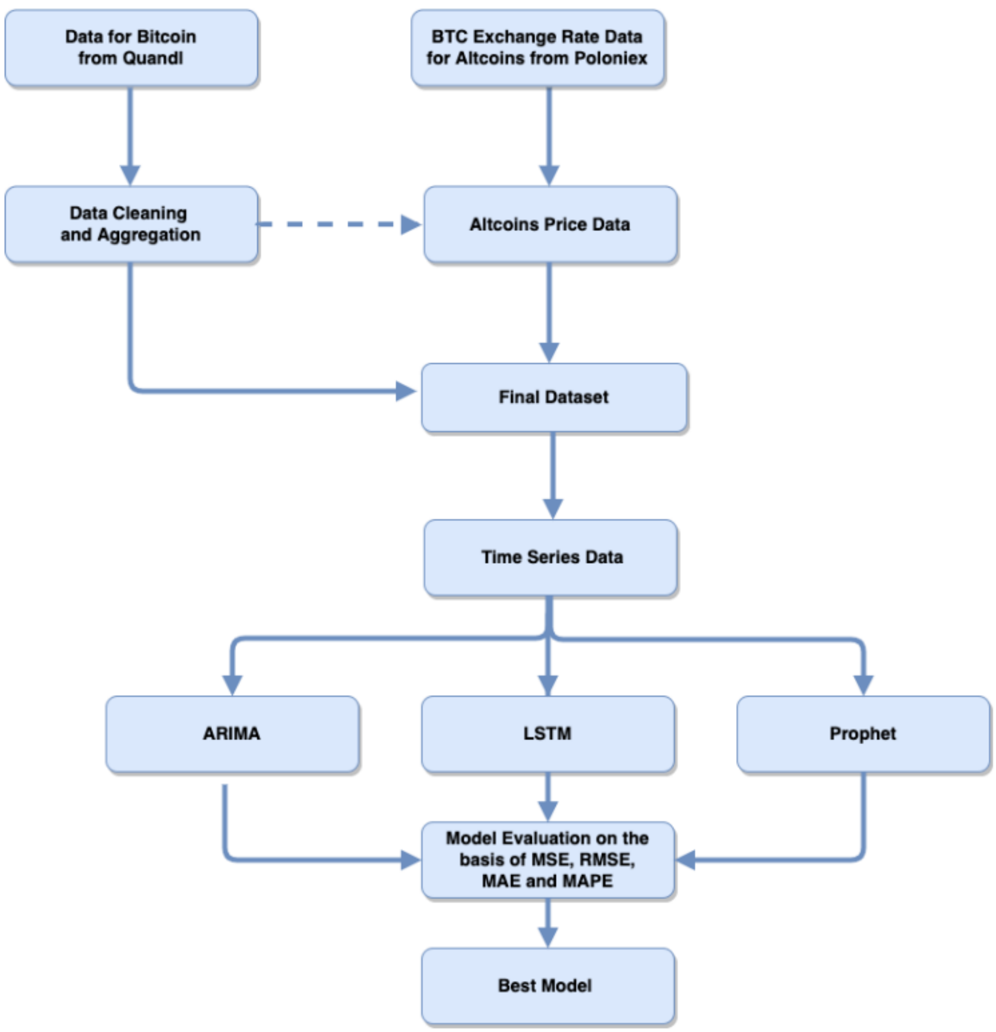


Figure 4.1 – Implementation Process in a Flow

A lot of datasets are available for bitcoin and other cryptocurrencies from different sources. The Quandl API can be used to extract Bitcoin data from Quandl website. The pickle library can serialize and save the data to provide data frame with all gathered data. The data from “Kraken bitcoin exchange” was visualized and retrieved to look for any discrepancies. There was a presence of only a few zero values. It is not that simple to figure out the exact bitcoin prices as demand and supply should be considered to determine prices. Coinbase, Bitstamp and Itbit are three of the key Bitcoin exchanges.

Out of these exchanges, price data was gathered and it is possible to combine all four values of pricing into an individual data frame as per “Weighted Price”. Some inconsistencies have been seen after visualizing the data and prices are mostly close to one another. To avoid any zero values, data is cleaned and average of all the prices are estimated to find out the final price for examination. “Figure 4” illustrates the flow of data aggregation and cleaning of “bitcoin price data” clearly where the top left graph illustrates the Kraken prices and top right graph represents all four exchanges’ prices. On the bottom left, the graph is achieved after aggregation and data cleaning, while the bottom left graph presents average prices of bitcoin over the years.

The Poloniex API was applicable for gathering data for other cryptocurrencies and for each altcoin, it was possible to extract Bitcoin exchange rate. The altcoin prices were determined in USD with previous Bitcoin data and existing data by multiplying average BTC price and altcoins’ weighted price. Finally, a data frame of all prices was made to analyze the data. After aggregation and cleaning of bitcoin price data, the Litecoin and Ethereum price data was extracted and determined to find out if there are any null values. This data was gathered as Bitcoin exchange rate. This way, Bitcoin average prices are estimated to multiply the exchange rates to determine the USD prices for Ethereum, altcoins, and Litecoin.

Once data from cryptocurrencies are gathered, researchers have done exploratory data analysis for better understanding. From “statsmodels” Python module, SARIMAX and ARIMA libraries were imported. The analysis of price data was done to look for any seasonalities and trends. Bitcoin data showed changes over a time period in mean and variance. An ideal timeseries dataset should be stationary by nature and it is important to take care of the presence of broad trends. There is no stationarity in results of Dickey-Fuller test and “Seasonal Decomposition of data” was conducted by (Chaudhari, 2020) as illustrated in Figure 4.3.

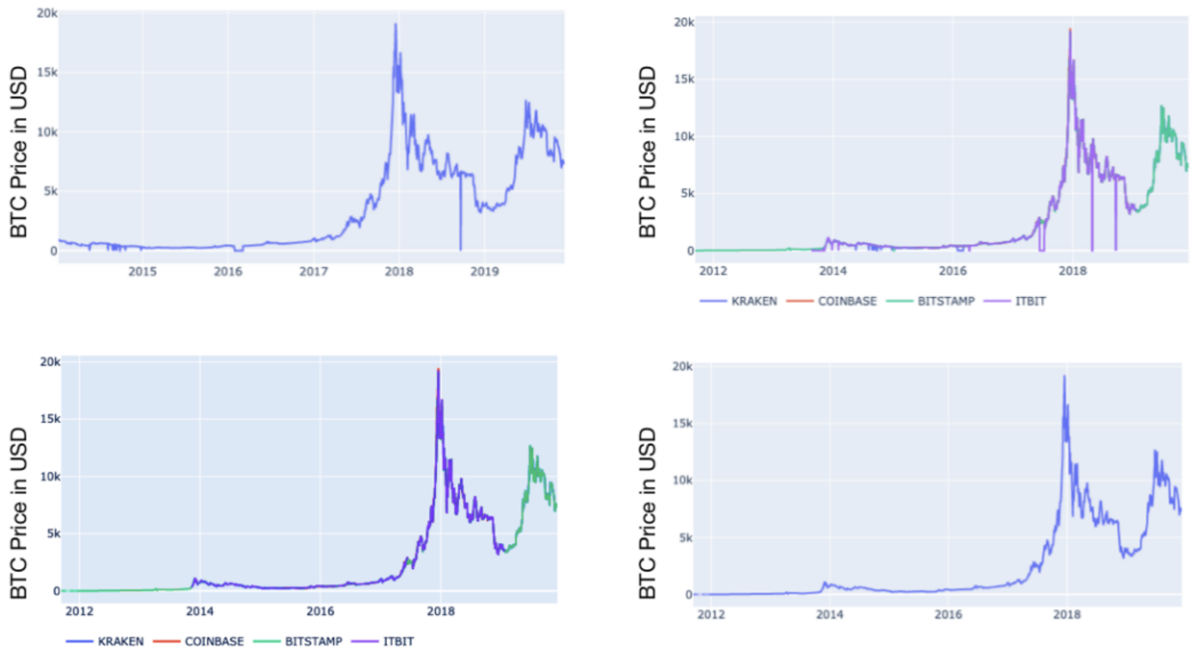


Figure 4.2. Aggregation and Cleaning of BTC Price Data

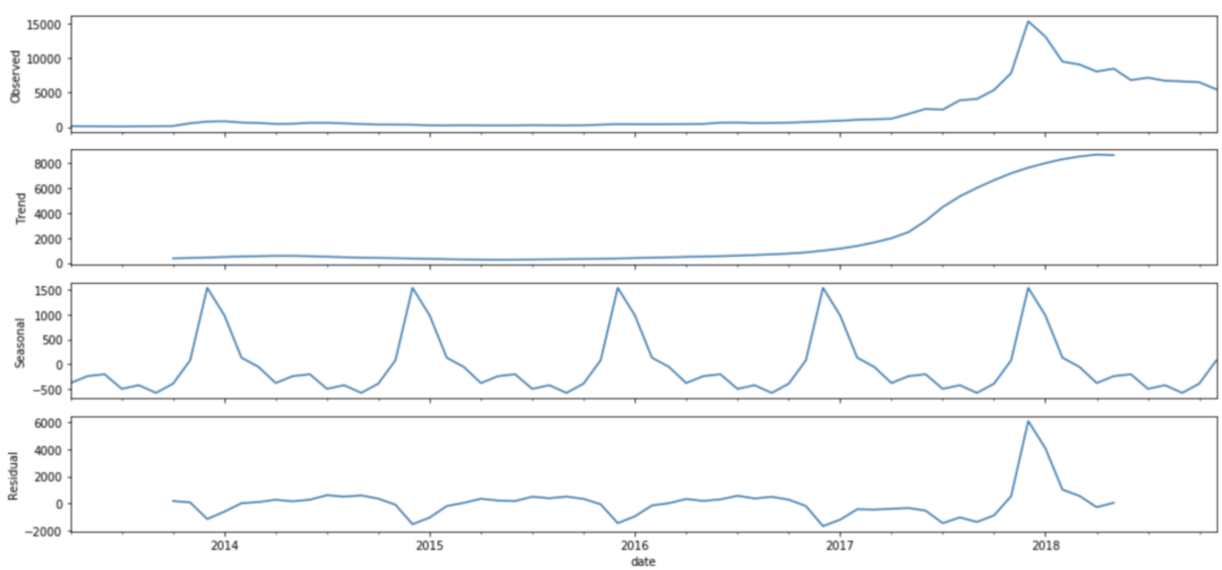


Figure 4.3. Seasonal Decomposition of data about BTC values

Then, they conducted “Box-Cox transformations” to suppress the variance after regular and seasonal differentiation. The autocorrelation shows the strength of a specific datapoint which is observed with datapoint that was previously observed. This way, partial autocorrelation and autocorrelation factors are made to check data patterns of given time series. Figure 6 shows how to observe positive correlation in first three to four lags. Majority of values are supposed to be in “blue region”, i.e. the “insignificant zone”. This way, there is a chance for seasonal components in the residuals. The combination of several parameters apply the ARIMA model and its quality assessment is fixed for each combination with “SARIMAX()”.

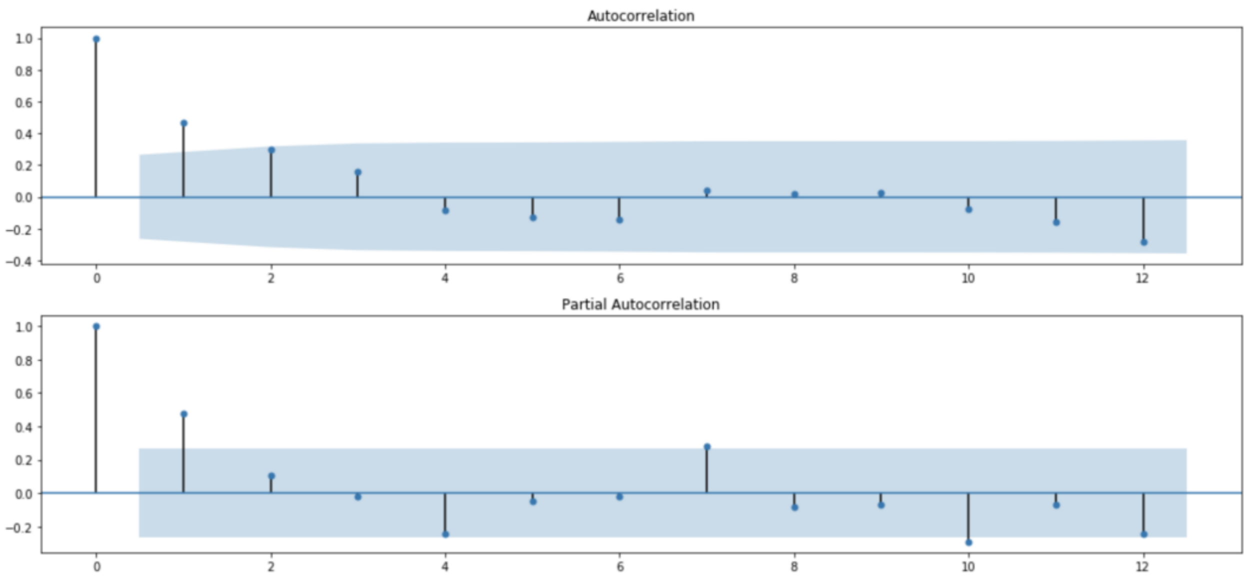


Figure 4.4. Autocorrelation and Partial Autocorrelation of Bitcoin values

Parameters are approximated initially along with the best model summary and selection of model to examine the diagnostics. It was assured that normal distribution was prevalent for residuals and there was a lack of seasonality. The Bitcoin price was forecasted and forecasting and prediction was possible with “invboxcox()”. The “Evaluation Section” discusses and evaluates the forecasted values. The stationary is checked with the same procedure for Litecoin and Ethereum price data. Keras is helpful to import all dependencies and LSTM library for LSTM model. Date is turned into datetime after loading data with pandas and split sequence is used to build the model.

Data is split into testing and training sets and timesteps have been assigned. The adam optimizer is used with bidirectional model for LSTM and training of data is done. Then, the bidirectional model is used for forecasting and evaluate the outcomes for all cryptocurrencies one by one. The fbprophet library is used to implement the Prophet model. Currency price and date columns are renamed to “y” and “ds” and data is loaded as it is. Column names should be in that way for the dataframe when using Facebook Prophet. Prophet model is applicable to data frame to forecast prices and price is turned into float by keeping the changepoint prior scale in mind along with “seasonality parameters”. It is followed by creating days in future to do prediction and predict prices. It can be done individually for all cryptocurrencies and one can visualize and evaluate forecasting results.

4.1. How efficiently machine learning models can predict cryptocurrency prices?

LSTM, ARIMA, and Prophet are the models implemented well and evaluated as per their “Root Mean Square Error (RMSE)”, “Mean Absolute Error (MAE)” and R^2 score values.

4.1.1. ARIMA

In ARIMA, the data was split into training set and test set data. The training set data was from January 2012 to December 2020, whereas the test set contained data from January 2021 to March 2021. For creating the ARIMA model, took exogenous features from the training data and created the model. It takes a lot of processing on the data to use it as time series for ARIMA as it is important to avoid seasonalities and trends in gathered data. After conducting a lot of statistical tests, the time series is conducted for “Auto ARIMAX” model and results were quite better in “RMSE” values. The R^2 score and MAE values shows that the model was ideal for their timeseries information. Table 4.1. shows the results using ARIMA –

Table 4.1. ARIMA Results

RMSE	**MAE**	**R^2**
2173.298	1741.95	0.95

Figure 4.5 illustrates the prediction of bitcoin prices using ARIMA model and lag is found between forecasted and observed values of all cryptocurrencies. The “Auto ARIMAX model” appears to work really well to predict “bitcoin price” till the last date.

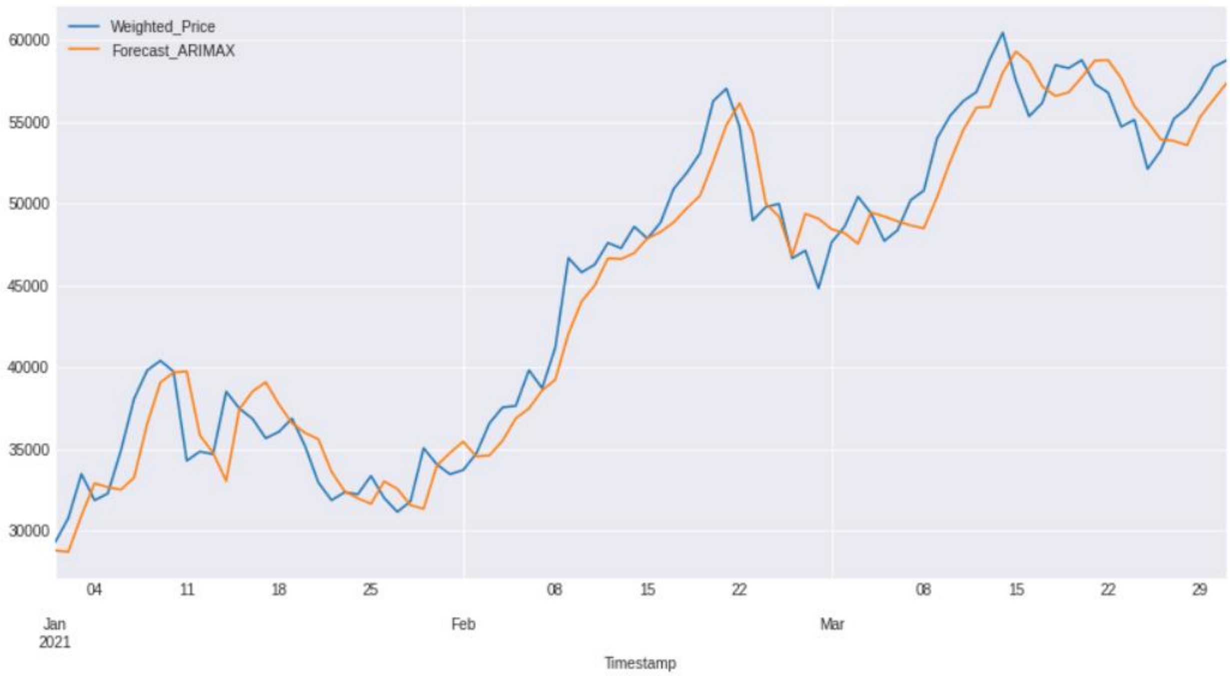


Figure 4.5. Forecasting using ARIMA model

4.1.2. LSTM

Table 4.2 illustrates the results achieved for cryptocurrencies with LSTM. In LSTM, for scaling the data, used MinMax scaler for this purpose. Then split the data into train and test respectively. Fit the RNN model with a batch size of 32, epochs as 50 and validation split as 0.1 values respectively. Once the model was trained, it was finally created on the test data set and checked the RMSE, MAE and R^2 score values. The result seems very promising as MAE and RMSE values are near perfect. RMSE value and the MAE values are around 0.3017 and 0.4414 respectively, proves that it might be the best model to forecast bitcoin prices. This proves LSTM an ideal model.

Table 4.2. LSTM Results

RMSE	MAE	R^2
0.3017	0.4414	-0.3973

Figure 4.6 and Figure 4.7 present price prediction with LSTM model and there no lag which can observed in predicted and actual prices of all the currencies. This model is better than ARIMA model.

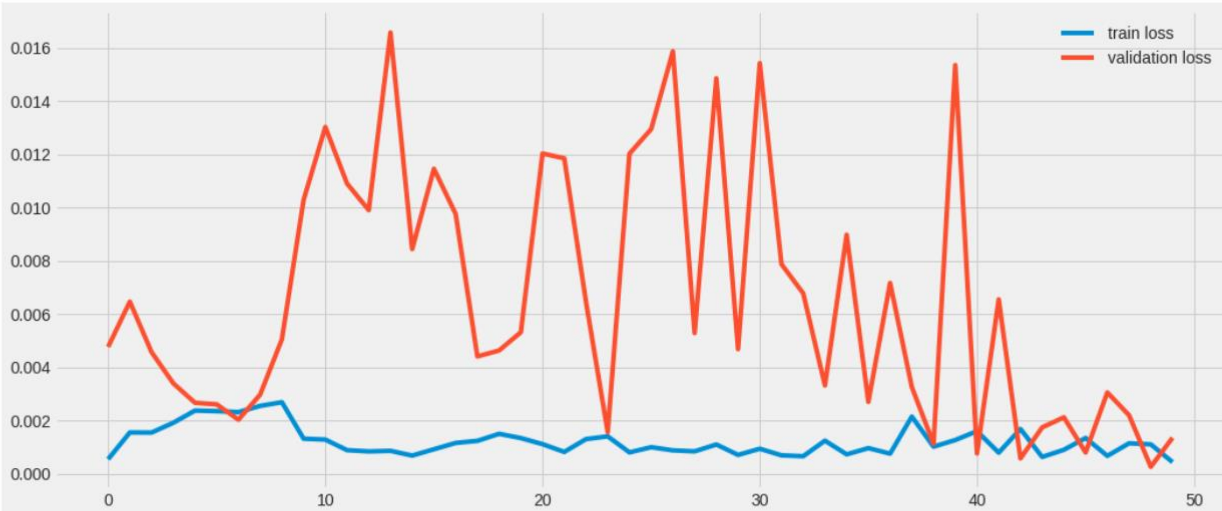


Figure 4.6. Plots of train loss & validation loss for Time Series Forecasting using LSTM

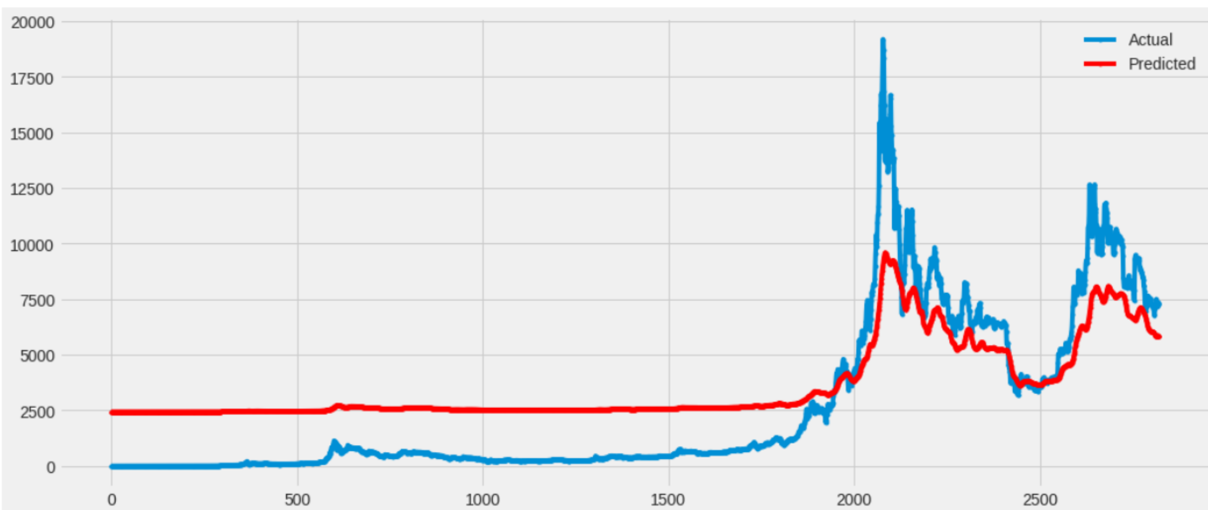


Figure 4.7. Plot of Actual & Predicted values of cryptocurrencies on test dataset using LSTM

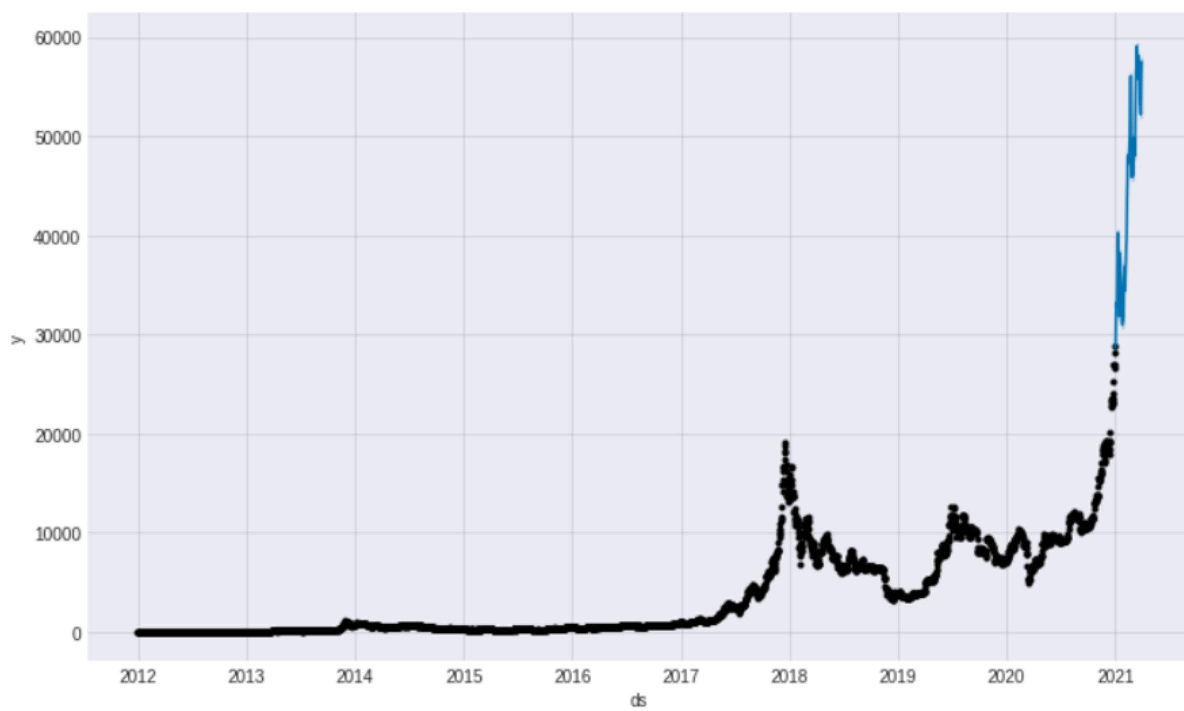
4.1.3. FB Prophet

In FB Prophet, resampling of the data to day level was done and a forward fill was done to handle the missing values. The train data was from January 2012 to December 2020 and the test data was from January 2021 to March 2021. The modelling was done using the same exogenous features. Once the model was trained, a final model was also made on the test data set as well. This model calls for least processing needed as it needs only the names of column as “y” and “ds”. Table 4.3 presents the results for the Prophet model. For Prophet model, the R^2 values achieved also seem well as values are above 0.94. But they achieved very similar MAE values and RMSE values to that of ARIMA model. The MAE value for the prophet model is around 1809 and the RMSE value is around 2273.

Table 4.3. – Results of Prophet Model

RMSE	**MAE**	R^2
2273.3109	1809.0152	0.9423

Figure 4.8 illustrates the plots of time-series forecasting using Prophet model. The blue line depicts the predicted forecast/the predicted values and black dots represent actual data points in our dataset. The plot shows upper and lower bound for prices which have been forecasted and it seems interesting. It is worth noting that forecasted values are synced with actual price mostly but it usually goes out of range at majority of datapoints. With significantly stable prices, fluctuations are shown along with more inconsistencies initially like just after the predictions of downfall in price which never occurred in real.




3675
3670
3665
3660
3655
3650
3645
trend
2021-01-09
2021-01-23
2021-02-06
2021-02-20
2021-03-06
2021-03-20
2021-04-03
ds
10
5
0
−5
−10
weekly
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week
200
100
0
−100
yearly
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year
55000
50000
45000
40000
35000
30000
25000
extra_regressors_additive
2021-01-09
2021-01-23
2021-02-06
2021-02-20
2021-03-06
2021-03-20
2021-04-03
ds

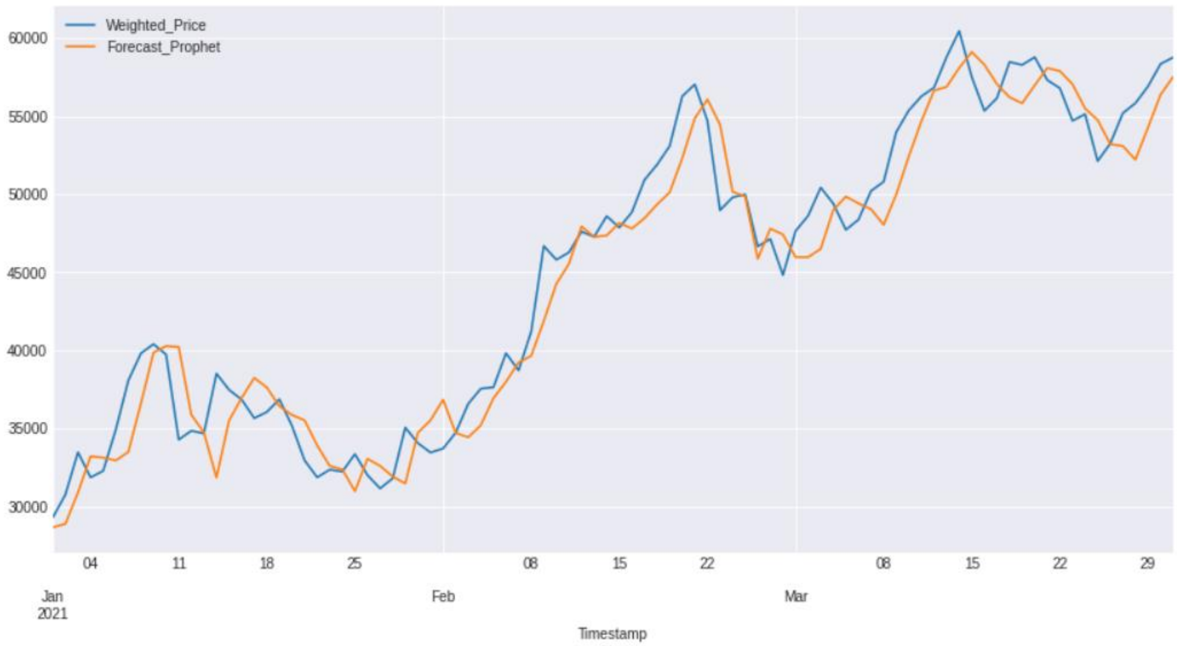


Figure 4.8. - Plots for Time Series Forecasting using Prophet

The "daily log returns" are the dependent variables for each crypto asset, which are calculated with their signs or closing prices. There are 50 variables which have set the overall input and most of them came from "raw data" after little bit of alteration. The log returns lagged around 1 to 7 days before off three cryptocurrencies in this set. These cryptocurrencies have returns which are interdependent highly at various frequencies "(Omane-Adjepong and Alagidede, 2019; Bação et al., 2018; Hyun et al. 2019)". For daily volatility, "RR_t (relative price range)" and σt "range volatility estimator" by "Parkinson (1980)" are the proxies calculated as –

$$RR_t = 2\frac{H_t - L_t}{H_t + L_t}, \tag{1}$$

$$\sigma_t = \sqrt{\frac{(\ln (H_t/L_t))^2}{4 \ln (2)}}, \tag{2}$$

The highest and lowest prices are H_t and L_t recorded at "t" day. To be specific, the set consists of first RR_t lag and 1 to 7 lags of σt for other Parkinson's applications to cryptocurrencies (Koutmos, 2018; Sebastião et al, 2017).

The input set includes the first set of network data and trading data of other exchange in the conforming cryptocurrency, excluding when they are unable to refute the null hypothesis of "augmented Dickey-Fuller (ADF)" test where the first lagged difference was used from the variable. There are seven variables on which this "differentiating transformation" was performed. The dataset also consists of seven dummies of deterministic day as it is observed that bitcoin and other crypto price dynamics may rely on the day (Aharon & Qadan, 2019; Dorfeitner and Lung, 2018; Caporale and Plastun, 2019).

In the "validation and training sub-samples", there is a "positive mean return" in all cryptocurrencies and "test sample" has a negative mean return. Only the Ethereum and bitcoin mean returns in the models are important at the levels of 1% and 5%, respectively. From 15th August, 2015 to 3rd March 2019, 0.21%, 0.33% and 0.19% are the daily mean returns for BTC, Ether and Litecoin, respectively, during the sampling period, which are significant at 10 percent. In all these subsamples and cryptocurrencies, the "median returns" are pretty different. The medium returns of Bitcoin are confident in three "subsamples", the median return of Ethereum is positive only in "validation subsample" (causing -0.12% as median return in the given period), while "Litecoin's median return" is negative over the past two subsamples, and 0 median return in the first subsample.

As discussed, there is a high volatility in these cryptocurrencies. It is evident from very high range length and standard deviations. For bitcoin, "standard deviations" start from 3.11% during the "training period" and Litecoin's standard deviations range from 8.05% during the validation period. Hence, volatility is above average by a factor of 10, at least, for all crypto assets. All the maxima and minima get back to 2-digit returns in percentage, with fullness for Litecoin is between extreme values and resulting in -39.52% (51.04%) "minimum (maximum)". It is worth noting that dynamics of Ether's volatility is shrinking over three periods and varies from that of the rest of two crypto assets. The volatility rises from the first to second subsamples for Litecoin and bitcoin and then lowers, accessing values a bit higher than in training period sample. Bitcoin has least volatility as compared to other cryptocurrencies.

In the third and first subsamples, there is a negative skewness for bitcoin and Ethereum in the 3rd subsample. Negative skewness "-0.26" is observed only in bitcoin, while Litecoin's skewness reaches 1.26 value. Especially in the training subsample, there is additional kurtosis in all cryptocurrencies. All of the "daily autocorrelations" are "positive" in the initial and next subsamples and "negative correlations" in the last order. Only the "Ether's autocorrelation" assumes 6.88% value during the training sample and at 10% significant level. The "autocorrelation coefficients" are overall low for BTC, ETH, and Litecoin at 0.41%, 4.18% and 1.31% respectively. It means that there is a lack of significant data in the daily returns to preview the next day returns linearly most of the time.

Classification or regression trees are combined in Random Forests (RF). Regression random forests are used in this application for next "return forecasting" and "classification RFs" can be used when it comes to achieve a "binary signal" for prediction of rise or decline of the price next day. The classification or regression tree is the basic block of random forests. It is a typical model which relies on recursive partition of space which is given by separate variables in small areas. Hence, the tree is read from the root or "first node" while predicting the value. In addition, successive branches are selected and tests are made until reaching the leaf node or terminal node, which defines predictable value for the dependent variable. Whether the price goes up or down next day is predicted by the binary signal or next return forecast. Several trees have been used by the Random Forest. A random subset in each "tree node" of "autonomous variables" and that of comments are used in "training dataset" to define the test which is helpful in branch selection. The forecasts of Random Forests are achieved on the average of forecasts made by various trees in regression RF or by selecting the binary signal with more trees in classification random forests.

"Support Vector Machines (SVMs)" are also useful for regression and classification tasks. SVMs attempt to look for hyperplane separating two outputs leaving largest margin in binary classification It refers to the addition of closeness to the "data point" (Yu and Kim, 2012). Slack variables may be allowable and introduced by classification errors measuring misclassification degree and parameter calculating the trade-off between amount of error and margin size.

The key here is to look for the model estimating output values to avoid larger errors than the given value (ε) in regression SVMs. It refers to the use of "ε-insensitive loss function" by the SVMs which neglects smaller errors and penalizes greater errors than this value (ε). A function is minimized for reaching the model coefficients which has a total of a reciprocal function of "margin" with penalty for larger deviations than ε between the original and predicted output values. Here is the function to estimate the SVM regression if {($x1$,$y1$),…., (xl, yl)} is the normalized, scaled, and centered training data –

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b, \tag{3}$$

Here, the inner product is (w,x). The "flatness" of the boundary can be seen in regression SVM of the given model "(Smola and Schölkopf 2004)", i.e., the joint of the "Euclidean norm of the vector of the coefficients", 1/||w|| "(Yu and Kim, 2012)". The parameter defining the "trade-of" between "prediction errors" above ε is denoted by C > 0. The slack variables determining those errors by "ξi and ξ_i^*" may estimate "Model (3)" by solving this equation by Smola and Schölkopf (2004) -

$$\min \frac{1}{2} \| \boldsymbol{w} \|^2 + C \sum_{i=1}^{l} \left(\xi_i + \xi_i^* \right)$$
$$\begin{cases} y_i - \langle \boldsymbol{w}, \boldsymbol{x_i} \rangle - b \le \varepsilon + \xi_i, & i \in \{1; \ldots; l\} \\ \langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + b - y_i \le \varepsilon + \xi_i^*, & i \in \{1; \ldots; l\} \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{4}$$

SVM can use "kernel trick" for handling non-linear models. First of all, it maps the original data into just another "high-dimensional space", where linear models can be applied for the issue. SVMs work on dual representation and this mapping relies on kernel functions. Support Vector Machines use linear models in this space of the data and also non-linear models in original space. In SVMs, some of the major kernel functions are the polynomial kernels and Gaussian kernels "(Ben-Hur and Weston, 2010; Tay and Cao, 2001)". "Gaussian kernels" are supposed to perform better with assumptions of general smoothness and they are used widely (Tay and Cao, 2001; Patel et al, 2015). Original linear models can also be used, which are usually known as "linear kernel". The RF packages are used to implement RFs and e1071 for implementing SVMs in R (Liaw and Weiner, 2002; Meyer et al, 2017). Torgo (2016) has also applied these approaches in real world.

In time-series machine learning applications the data are split widely in training set to determine a validation set and models where ideal model is selected and best models' results are calculated in the test set. Avoiding all data snooping risks and ensuring representative results to be achieved in test set are the main concerns when it comes to define various data subsets. Splitting the dataset into two equal subsample lengths is the first approach here. The first subsample is used only for building the first model for training by setting the data parameters. Then, the next part of the data is split into "25% of validation subsample data" and into the last "25% of testing subsample data". Then, the best model is selected using the "validation subsample" for each class and the profitability and forecasting performance of models is measured with the test subsample. A set of "hyperparameters" is defined for each period to pick the best algorithm and a set of "explanatory variables" for the same method.

The model assembling or averaging are pretty simple for basic machine learning models. There are several intricate assembling procedures but they are more likely to give better results. For example, this classification issue can be addressed by Kou et al (2012) for algorithm selection as the issue of "multiple criteria decision-making (MCDM)" and proposed to fix differences among approaches on the basis of "Spearman's rank correlation coefficient". The MCDM approach is introduced by Kou et al (2014) to rank the choice of major clustering models in "financial risk analysis" domain. Feature engineering is used by Li et al (2020) for performance improvements of classifiers to detect malicious URLs and "adaptive hyper sphere (AdaHS)" is proposed by Li et al (2017) as "adaptive incremental classifier" and "Nys-AdaHS)", its kernelled variant for dynamic data where patterns are subject to changes.

A "performance indicator" battery is used to assess whether trading strategies are profitable. The win ratio is similar to the day's ratio when the right positive sign is given by the "ensemble model" for the total number of the days in market and the next day. It also shows the SD and Mean of the returns in case of active positions. The year-wise "compound return" is the "annual return" as given by the combined daily and "discrete returns" with all days given in the "test sample", along with "zero-return days" when there are no plans given in the market.

The “ratio between SD of daily returns and daily return” refers to the annualized “Sharpe ratio” with all test sample days multiplied by “√365”. The bootstrap p-values are the odds of “daily mean return” of the model which has been proposed and is above the “daily mean return” considering all days in sample, given the null for having equal mean returns. The “Conditional Value at Risk (CVaR)” measures the tail risk at 1% along with the “maximum drawdown”. The “average loss” provisional on the “Value at Risk (VaR)” is measured at 1% level by the former. The “maximum loss” observed is measured by the latter from the top to a manger of the “combined value of trading” pattern before attaining a new peak, associated with the value of the same peak. As given in several exchange markets by Alessandretti et al (2018), the proportional costs of transaction are usually ranging from 0.1% to 0.5%. Even the “high-fee online exchange” is traded by the investor, it is observed that 0.5% of “round-trip transaction” is ideal estimate of the total trading cost, along with implicit and explicit costs like price impacts and bid-ask spreads. This figure is higher than that in various related studies.

4.1.4. XG Boost

Table 4.4 illustrates the results achieved for cryptocurrencies with XG Boost. In XG Boost, have used hold-out based validation method. For modelling purpose, used Random Search CV for best output. The result does not seem good at all as MAE and RMSE values are very high. RMSE value and the MAE values are around 17843.1253 and 20117.8111 respectively, proves that it might not be the best model to forecast bitcoin prices.

Table 4.4. – Results of XG Boost Model

RMSE	MAE	R^2
20117.8111	17843.1253	-3.5154

Figure 4.9 illustrates the plots of time-series forecasting using XG Boost model. The blue line depicts the Weighted Price values and red line represents the forecasted XG Boost values in our dataset. The plot shows upper and lower bound for prices which have been forecasted and it does not seem interesting at all. It is worth noting that forecasted values have not synced with actual price mostly but it usually goes out of range at majority of data points. With significantly stable prices, fluctuations are shown in the Weighted Price along with more inconsistencies initially and

then it gradually increases whereas, the forecasted values have been stable throughout the due course of time.

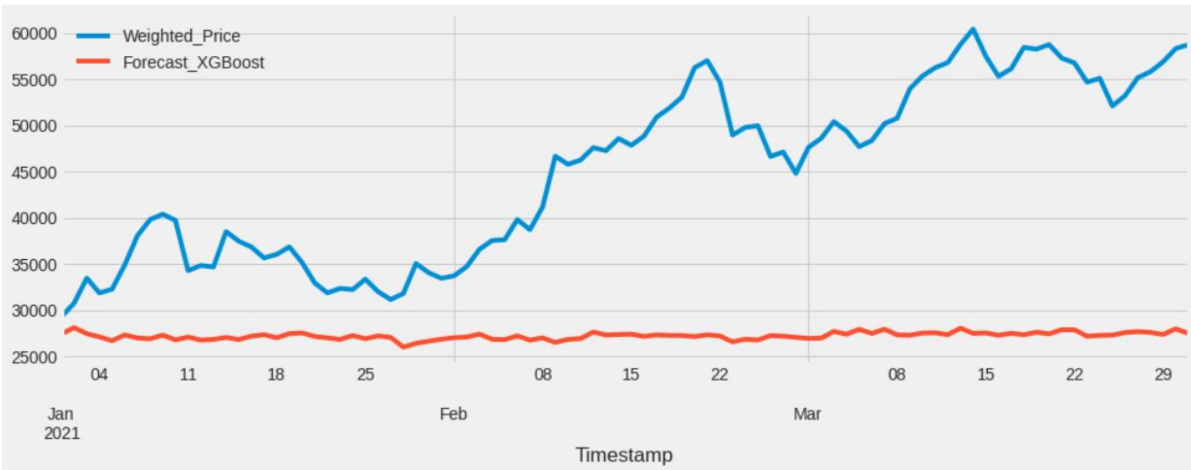


Figure 4.9. - Plots for Time Series Forecasting using XG Boost

4.2. Implementation

This section discusses implementation of machine learning methods on earlier studies for cryptocurrency time series forecasting and concludes the output.

Derbentsev et al (2019) implemented "short-term prediction model" for time-series forecasting. They adapted the up-to-date "Binary Auto Regressive Tree (BART)" to standard and time-series models. BART is found more accurate as compared to ARIMA in transitional and slow-growing dynamic times.

Yiying & Yeze (2019) used cutting-edge AI frameworks for the well-connected "Long Short-Term Memory (LSTM)" and "Artificial Neural Network (ANN)" networks to analyze dynamics of Ethereum, Bitcoin, and Ripple. It is found that longer-term history is used by ANN while more "fast dynamics" were used by LSTM, which shows greatness of LSTM as compared to effectiveness of ANN when it comes to use historical data.

Lahmiri and Bekiros (2019) used educational technologies to determine prices of Bitcoin, Ripple, and Digital Cash, three of the most commonly traded digital currencies. They used it for the first time as per their experience in cryptocurrency experience. Even though computing burden was higher in LSTM model as compared to “nonlinear pattern brutality”, profound learning was very important to predict “volatile dynamics” in crypto markets (Hu et al, 2021; Sen et al, 2020; Sen and Chaudhuri, 2018). Researchers have also implemented various DL and ML algorithms to assess and forecast factors driving crypto prices (Patel et al, 2020; Chen and Sun, 2020), such as “Long Short Term Models (LSTM)”, “Neural Networks (NN)”, and “Gated Recurrent Unit (GRU)” (Dutta et al, 2020). Iqbal et al. (2021) implemented a “hybrid cryptocurrency LSTM” prediction model on only Montero and Litecoin. The proposed model was found reliable in predicting values and scheme is viable for various forecasts.

Bitcoin prices are identified with high frequency and regular prices by Chen et al (2020) for prediction of bitcoin prices using ML techniques at varied frequencies. SDA and XG Boost have been found with higher results than common benchmarks with accuracy of 65.3% and 66% respectively. Rathan et al (2019) extracted and compared accuracy of prediction of bitcoin prices with several ML models and found correlation between regression model and decision tree in experiments.

Borges and Neves (2020) proposed an investment strategy to trade cryptocurrencies in exchange markets. They explained all learning models, whether resampling or not, while dealing with B&H strategies in over 100 markets which have been surveyed. The best overall performance is recorded in unweighted average from learning models, i.e. around 59.26% of accuracy in time sampling. Both sampling methods have been tested to provide very high returns and smaller probability of retrieving data on timely basis. Kwon et al (2019) used LSTM model structure and compared f1 values. The time series range is exceeded by the LSTM model. The “time series price range booster” is the master learning model with relatively better prediction. Sebastião and Godinho (2021) assumed ensemble learning to give the best output of “assemblies 5” or 5 Comparable Signal models with 80.17% annualized ratio and Sharpe with 91.35%, apart from 9.62 and 5.73 percent of annualized returns, respectively. It is argued that ML provides robust approaches for time-series forecasting and creating the right strategies even in extreme situations in positive findings.

Various feature selection models have been tested by Mallqui and Fernandes (2019) for the best prediction attributes in initial technique. They tested the behavior of SVM, ANN and Ensemble models for price trend prediction, followed by clustering for k-means and RNN. SVM and ANN are also viable to return closing, minimum, and maximum prices of Bitcoin. Along with it, the researchers also used regression results as inputs to increase the forecasts of price route. They found improvements in best machine learning model and selected attributes.

Poongodi et al (2020) forecasted concentric pricing with two ML models, SVM and LR, and common market closing series to predict the pricing of Ether. They used various weight coefficients and window lengths of filters in prediction of Ether prices. They used cross-validation to develop best-performing model, despite the dataset, in training model. The proposed model is implemented with two machine learning algorithms. They found more accuracy in SVM (96.06%) than LR (85.46%) after applying the proposed model. In addition, it is possible to increase accuracy of the given model to 99% by adding features to SVM.

CHAPTER 5

RESULTS

5.1. Introduction

The data analyzed above presented the forecasted price for cryptocurrencies with four algorithms, i.e. a statistical model ARIMA, a DL model LSTM, a time-series forecasting algorithm Prophet and XG Boost model. All these models are different by nature and each of them use different techniques for price prediction. Table 5.1 summarizes the results as per R^2 values and it is worth noting that LSTM performs better than Prophet, ARIMA and XG Boost as per correlation between forecasted and observed values. MAE values showed weird results for XG Boost and MAE values should be optimized or improved.

Table 5.1. Results of R^2 for four models

R^2	LSTM	ARIMA	Prophet	XG Boost
	-0.3973	0.9473	0.9423	-3.5154

These are not extraordinary results because of two common reasons. First reason is limitation. Only historical crypto prices are taken in this study for time series forecasting, rather than other factors which might affect their prices. It is worth incorporating several features which may fluctuate crypto prices. Other factors should also be included to forecast prices for small time if forecasts are used for trading instead of long periods as it has scope to learning models and to minimize the error significantly.

The rise and fall of prices are also responsible for high error values over the past two years. Most of the studies referenced and reviewed have used cryptocurrency prices by 2018 and they don't have another major hike in historical price of bitcoin in 2019. The major variation in crypto prices and trading volumes in altcoins and top tier is another factor. More or less, these cryptocurrencies show the same trajectory over the past couple of years. Hence, there is no major sign of contradiction in the results in efficiency of models. If a different band of altcoins is selected, which may be either very stable or volatile, results may differ.

Machine learning refers to studying algorithms which improve on their own with experience and by using data. It is supposed to be the part of artificial intelligence. A model is built by ML algorithms which supported sample data or “training data” in order to predict without being programmed explicitly to do so. On the basis of various models or methods used, the training of the models was done on previous information on bitcoin value from December 16, 2020 to December 29, 2020 and predictions and results are compared by all machine learning models from December 29, 2020 to December 31, 2020. It is because there was a steep hike in bitcoin prices during this period, i.e. whopping US$8000 approx.

5.2. Comparison of RMSE and MAE

In order to choose the best model in worst case scenarios, extreme volatility of bitcoin values is also considered. For each model, “Root Mean Square Error (RMSE)” and “Mean Average Error (MAE)” are two major performance metrics computed (Willmott & Matsuura, 2005). Table 6 provides a detailed overview of those results in which Wadalkar et al (2021) compared the performance of each model with other models for prediction, apart from giving the reason to reject or choose a specific model for serving purpose.

Table 5.2. Comparing RMSEs and MAEs of various models

Models	**RMSE**	**MAE**
XG Boost	20117.8111	17843.1253
LSTM	0.3017	0.4414
Prophet	2273.3109	1809.0152
ARIMA	2173.2980	1741.9507

5.3. Analyzing ML Models

- **LSTM** – Figure 5.1 illustrates forecast vs. real prices of LSTM model from January 01, 2012 to March 31, 2021. The “Long Short-Term Memory” model is trained by the “sigmoid activation” feature for 50 epochs with Adam optimizer. The error between actual and predictions is compared via “Mean Average Error (MAE)” and “Root Mean Squared Error (RMSE)”. For LSTM, the MAE is 0.4414 and RMSE is 0.3017, which is the lowest in comparison to other models. This is the best model with utmost accuracy.

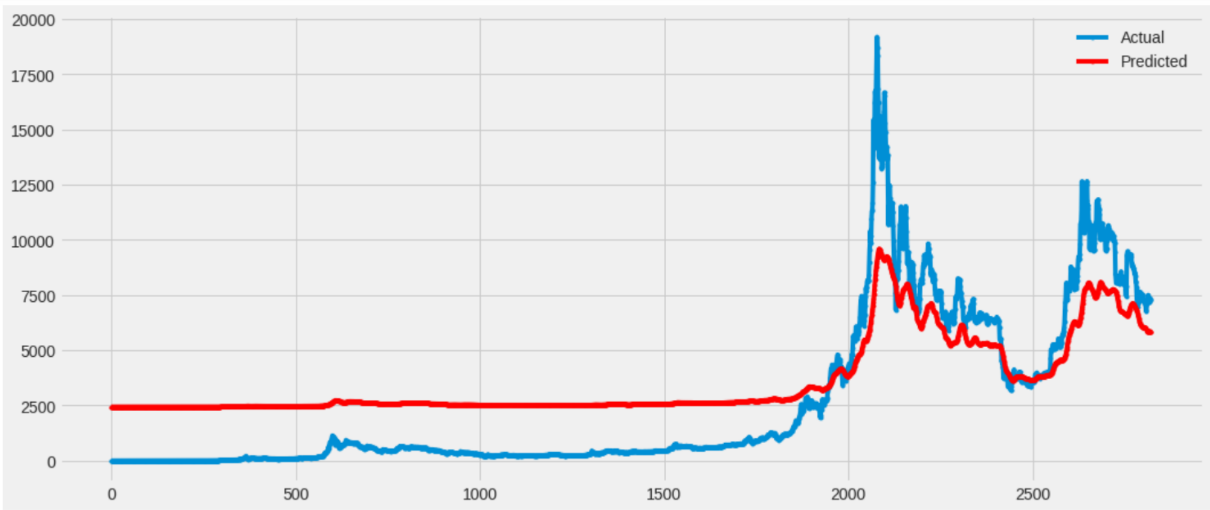


Figure 5.1. – Comparison between Actual and Predicted values for LSTM

- **XG Boost** – Figure 5.2 illustrates XG Boost Actual and Forecast values in the plot during the period of January 01, 2012 to March 31, 2021. For XG Boost, the RMSE is 20117.8111 and MAE is 17843.1253, which is not accurate in comparison to other models. Hence, it is not much accurate.

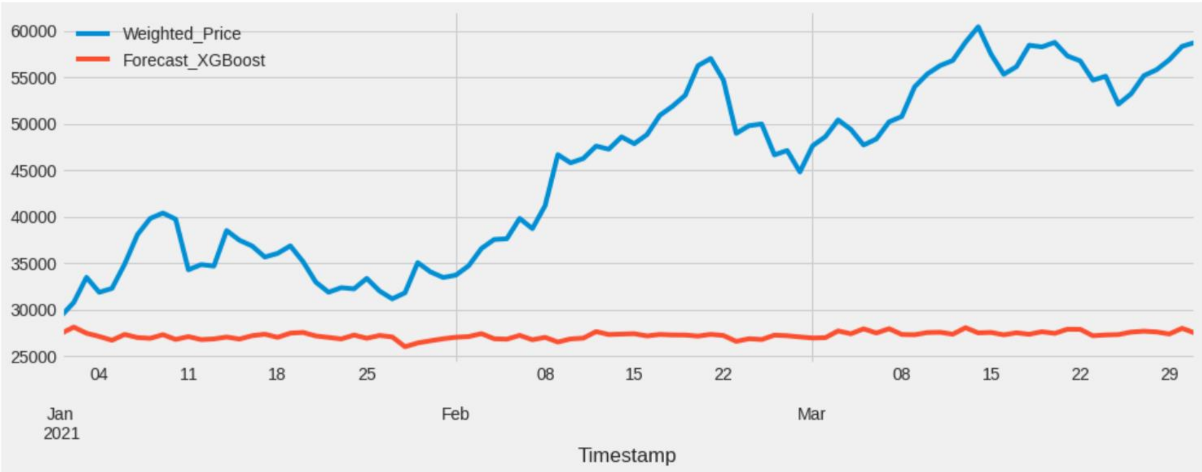


Figure 5.2. Comparison between Actual and Forecast values for XG Boost

- **FB Prophet** – Figure 5.3 illustrates FB Prophet plot for Actual vs. Forecast values from January 01, 2012 to March 31, 2021. The MAE is 1809.0152 and RMSE is 2273.3109, which is less accurate in comparison to both XG Boost and LSTM. Hence, it is not that flexible for prediction of volatile data and not the best model either.

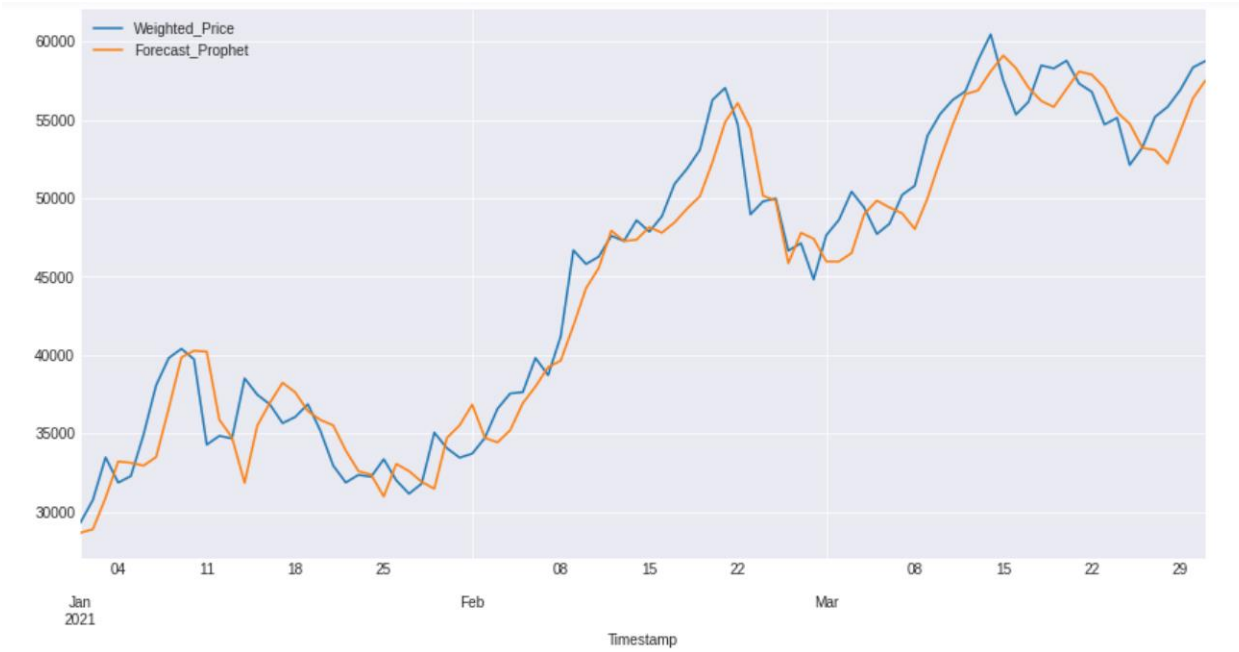


Figure 5.3. Comparison between Actual and Forecast values for FB Prophet

- **ARIMA** – The "Auto Regressive Integrated Moving Average" forecast for January 01, 2012 to March 31, 2021 is featured in Figure 5.4. For ARIMA, the MAE is 1741.9507 and RMSE is 2173.2980, i.e. high in comparison to above models. Hence, it is also not the best suited model.

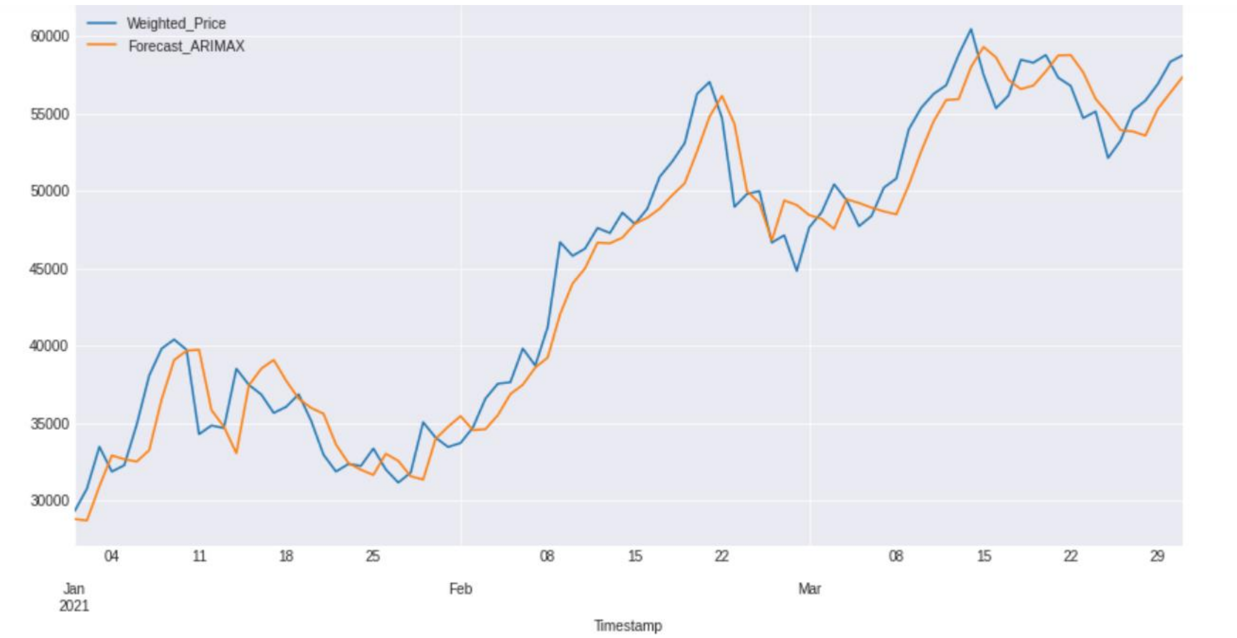


Figure 5.4. Comparison between Predicted and Average values with ARIMA

Due to global recession in 2008 and lack of trust with the financial system, bitcoin has become the prominent part of international financial market with huge attention of governments, regulators, individual and institutional investors, common public, vast media coverage, and academia since inception. In 2018, "what is Bitcoin" was the most searched term on Google US and UK platform (Marsh, 2018). In December 2017, the "Bitcoin futures" contracts were launched by the "Chicago Mercantile Exchange" and "Chicago Board Options Exchange (CBOE)" which is a sign that financial industry has always tried hard to stay off from this trend.

With price appreciation and significant growth of market capitalization, Bitcoin's success has led to the emergence of several other cryptocurrencies which are different mostly in a few parameters from bitcoin, such as currency supply, block time, and issuance scheme. The crypto market has been one of the leading unregulated markets worldwide by now, totaling over 5.7k cryptocurrencies, over US$270 billion of market capitalization and 23000 online exchanges (Foley et al, 2019). Satoshi Nakamoto initially designed Bitcoin as a P2P payment mode in 2008. Later on, other cryptocurrencies were developed. But bitcoin has achieved great reputation as a completely speculative asset.

Most of the time, their prices are idiosyncratic as they are driven majorly by behavioral elements and are uncorrelated with major financial asset classes but their efficiency of data is still the matter of debate. At the same time, a lot of asset managers and hedge funds have upgraded their portfolios with cryptocurrencies and a lot of studies have been done in crypto trading while focusing on ML models (Fang et al, 2020). This study opens further research path on determining profitability and predictability of various cryptocurrencies with machine learning models. Hence, this study also has contribution to this recent literary trend on cryptocurrencies. Researchers have chosen cryptocurrencies as per their common characteristics, age, trading volume, market capitalization, and media attention and Bitcoin, Ether, and Litecoin cover over 75% of overall market capitalization of all other cryptocurrencies.

As a P2P digital currency, bitcoin was successful in the beginning as it solves the issue of double-spending with its “crypto-based technology” which eliminates the requirement for trusted third party. Blockchain is an emerging technology for bitcoin as it works as a permissionless or public digital ledger which records transactions between the users. With the existence of central authority, participants or nodes can replicate this ledger of the network. The dedicated software is used to maintain the same collectively (Yaga et al, 2019). There are several features of “bitcoin ecosystem”. It is an online system on the basis of crypto entities without any intrinsic value of representation and it doesn’t require third-party or trusted intermediary. It is also open-source or consensual as the network transfers and balances bitcoin.

On August 2015 and October 2011, Ethereum and Litecoin were launched, respectively. The protocol is similar between bitcoin and Litecoin and has “84 million units” for supply. It was made to save “computation power” needed for mining in order to improve the speed of overall processing and to perform transactions quite faster, which is especially attractive in situations where time is needed. Ethereum is also a peer-to-peer network but it is a token unlike Litecoin and bitcoin. It is also known as Ether with no maximum threshold. In addition, it provides platform which allows applications on its blockchain in a way that it can be used by public as a decentralized ledger. To be specific, it facilitates the applications of smart contracts or “contractual agreement” with least possibility of censorship, downtime, third-party interference, or fraud. These features have gained the interest gathered by Ethereum since inception and it has become the next most vital cryptocurrency.

Whether bitcoin is completely a speculative crypto asset or just another currency type has been debated widely in early studies and majority of authors found it with high volatility, bubble-like price fluctuations, and returns in the very short run (Dwyer, 2015; Yermack, 2015; Cheah and Fry, 2015; Cheung et al, 2015). This claim has been moved to other crypto assets which have been implemented well, including Ripple, Ethereum, and Litecoin "(Catania et al, 2018; Gkillas and Katsiampa 2018; Corbet et al, 2018a; Catania et al, 2018; Charfeddine and Mauchi 2019)".

The claim that crypto assets are completely hypothetical with no "intrinsic value" has led to a trial on common relations with "financial and macroeconomic variables" and on the rest of factors in behavioral domain of investors. These factors have been very vital also for more common markets. For example, according to "Wen et al (2019)", Chinese giants with more emphasis from "retail investors" are more likely to have lower risk related to stock price crashes. Kristoufek (2013) discussed "high correlation" between "search queries on Wikipedia", "Google Trends", and "bitcoin values".

"Kristoufek (2015)" strengthens earlier results and still couldn't find any "significant correlation" with key variables like "gold price in Swiss francs" and "Financial Stress Index". Bouoiyour & Selmi (2015) studied the relation between various variables and bitcoin prices like Google searches, gold market price, and bitcoin velocity and found only significant impact with lagged Google searches at 1% level. According to "Polasik et al (2015)", the formation of "bitcoin price" is driven mainly by "news sentiment", "news volume", and "bitcoins traded". There are 21 common drivers of returns from bitcoin determined by Panagiotidis et al. (2018) and found that Google Trends' search intensity is among the most important determinants. Panagiotidis et al. (2019) found lower impact of intensity of online searches on BTC prices, while these prices have strong positive impact from gold shocks.

According to Ciaian et al (2016), major factors driving bitcoin prices are investor sentiments and market forces and there is no clue for any impact of macro-financial variables finally. According to “Zhu et al (2017)”, there are financial variables like “Dow Jones Industrial Average”, “Consumer Price Index (CPI)”, gold price, “federal funds rate”, and “US Dollar index” have a great influence on bitcoin prices every month. According to Li and Wang (2017), speculative investment drives bitcoin prices in early stages and bitcoin prices are also deviated with economic conditions. The price dynamics followed economic changes more closely as the market matured, such as interest rates, cash flow in the US, inflation and GDP.

According to Dastgir et al (2019), there is a two-dimensional causal relation between “Google Trends-based bitcoin attention” and its returns in distribution tails. Bitcoin is not correlated with conventional asset classes like bonds, stocks, commodities, and exchange rates both in cases of financial crisis and in normal times (Baur et al, 2018). Bouri et al (2017) documented a feeble relation between bitcoin and other important financial variables like bonds, major “world stock indices”, oil, gold, “US dollar index” and “general commodity index”. There is also no relation between announcements on “Producer Price Index”, “employment rate” and “CPI in the US”, according to Pyo and Lee (2019). According to their results, bitcoin reacts to monetary policy announcements on the “Federal Open Market Committee” in the US.

Public recognition is another major driver of bitcoin prices, which is measured by Google searches, views on Wikipedia, social media updates, comments on forums or Facebook, and Tweets (Li and Wang, 2017). User replies and comments in online communities to predict fluctuations in bitcoin, Ripple, and Ether transactions, daily prices, with positive returns are considered by Kim et al (2016). The hidden “Markov models” were used on the basis of online “social media indicators” to make ideal trading strategies on various cryptocurrencies by Phillips and Gorse (2017). According to “Corbet et al (2018b)”, “Ripple, Bitcoin, and Litecoin” are not associated with various financial and economic variables in frequency and time domains.

Factors like trading volume, market data, attractiveness, and volatility affect “weekly prices” of “Ethereum, Bitcoin, Litecoin, Dash, and Monero” (Sovbetov, 2018). “Phillips and Gorse (2018)” determined the dependence of prices of cryptocurrencies and social media and online factors on market regime and observed that “medium-term positive correlations” meaningfully reinforce throughout “bubble-like” events, while specific market events cause the appearance of “short term relationships” like security breaches or hacks.

The Bitcoin dataset consists of “Open High Low Close (OHLC)” data from January 1, 2012 to March 31, 2021. The prediction period was short due to very high volatility of bitcoin prices, i.e. from December 29, 2020 to December 31, 2020. ARIMA, XG Boost, LSTM and Facebook Prophet are the most common machine learning models used for timeseries forecasting. “Root Mean Squared Error (RMSE)” and “Mean Absolute Error (MAE)” are widely used parameters for this purpose. LSTM has been found to be ideal model among other three algorithms with RMSE of 0.3017 and MAE of 0.4414. The way LSTM performed better than other models shows that it considers different decompositions like “Sessional and Residual Trend” and “Regular “Trend”. Even with higher deviation of prices than normal, LSTM is found to be very accurate for predictions for both long-term and short-term prices.

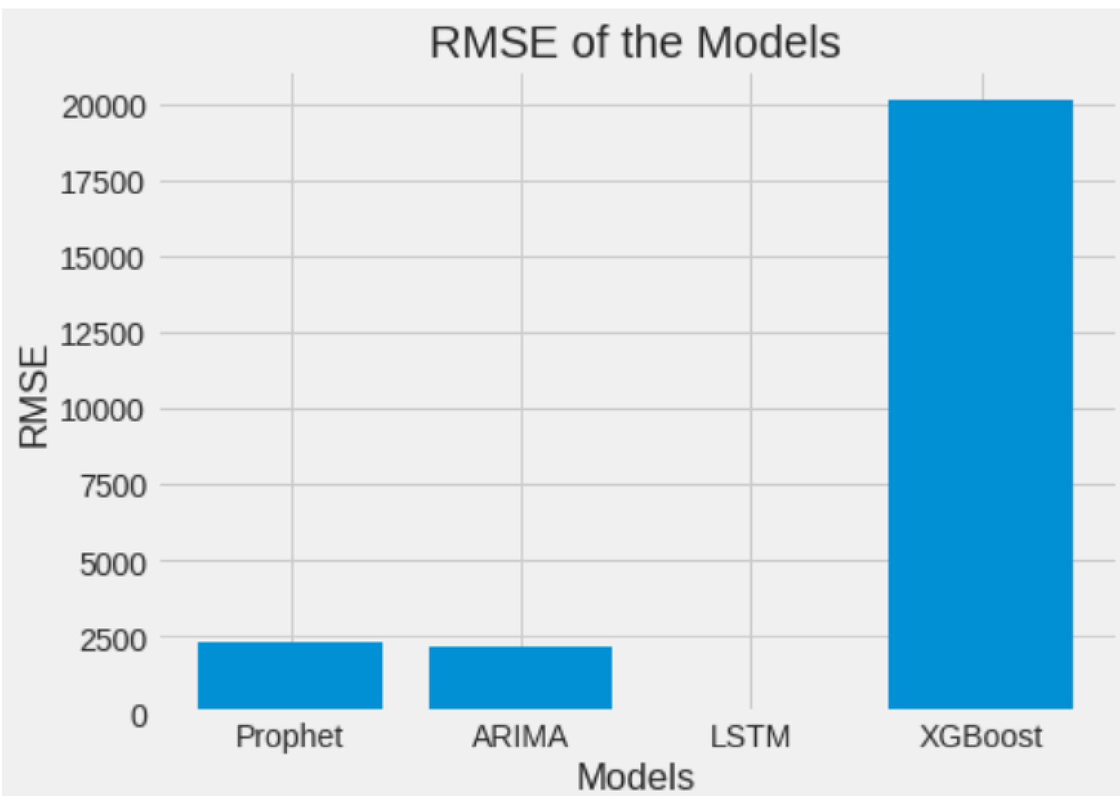


Figure 5.5. Comparison of RMSE values between different models

The hypothesis of "non-rational behavior" like herding is studied by researchers like Stavroyiannis and Babalos (2019) on "crypto market". "Gurdgiev and O'Loughlin (2020)" determined the relation between "price dynamics" of ten crypto assets and proxies for "uncertainty (US Equity Market Uncertainty Index)", "fear (VIX Index)", "investors' sentiments (as per their opinions on bitcoin forum)" and "investor perceptions" on whether financial markets would be bearish or bullish as per "CBOE put-call ratio". According to the researchers, investor sentiment must be considered to determine whether the crypto prices would go and investors must use cryptocurrencies as hedge for rainy days, but they are not ideal save haven at the events of fear against equities. The herding biases also play a vital role among crypto investors and found that recency and anchoring biases are "environment-specific" and non-linear, if present.

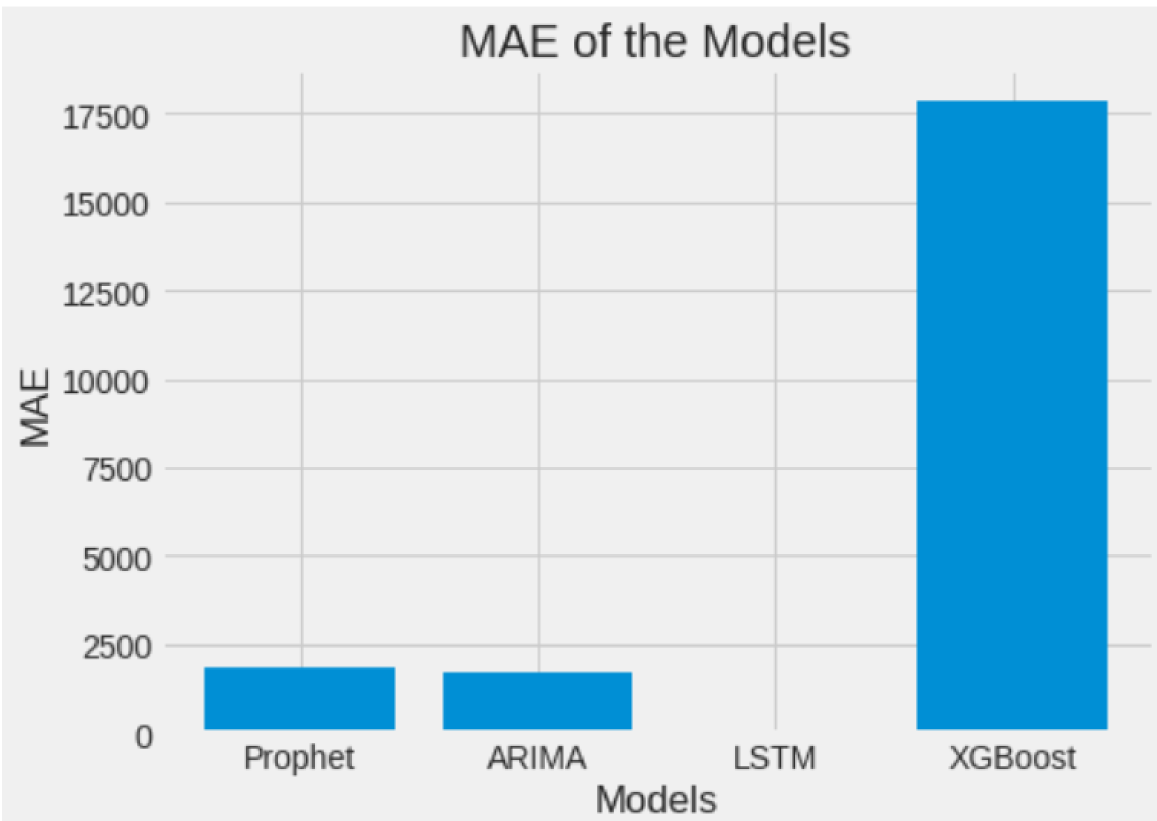


Figure 5.6. Comparison of MAE values between different models.

"Chen et al (2020a)" determined the effects of "fear sentiment" on BTC values and observed that high trading volume and negative returns were the byproducts of uncertainties related to COVID-19. Bitcoin worked more like the rest of financial assets rather than safe haven during the events of market crisis, such as COVID-19. Various authors have studied the efficiency of crypto markets like bitcoin. Using various approaches, Bariviera (2017) and Urquhart (2016) argued that bitcoin is not that efficient, while Tiwari et al. (2018) and Nadarajah and Chu (2017) had totally different opinions. However, Bariviera (2017) and Urquhart (2016) also argued that bitcoin has been efficient due to the first transition phase when market gets mature. The interest on machine learning techniques for profiting and forecasting from cryptocurrencies has been increased over the past few years.

5.4. Summary

All in all, machine learning models improved predictability of cryptocurrency returns and prices, presented great accuracy, outperformed various models like “Exponential Moving Average” and “Autoregressive Integrated Moving Averages” irrespective of period under investment horizon, analysis, input set, data frequency, method, and classification or regression type. Various studies computed the trading strategies’ performance as per “machine learning models” and against the passive “buy-and-hold (B&H) strategy with or without trading costs”. There is a lack of “unambiguous winner” when comparing various models. Overall, strategies derived on machine learning are better in volatility, “cumulative return” and “Sharpe ratio” as compared to passive strategy. Majority of researches focused only on bitcoin and covered “steady upward price trend” period and neglected “short-selling restrictions” and trading costs.

CHAPTER 6

CONCLUSION

6.1. Introduction

This study is based on forecasting the different Crptocurrencies and predicting their prices on the basis of historical data with various ML algorithms and achieved different results. The plots and values in one study, LSTM has shown the best results as compared to the other models when it comes to forecasting and predicting the cryptocurrency prices. However, these models are not restricted to price prediction only. They can also perform well for several predictions and forecasts and they could be used in different sectors along with financial or stock markets. This study suggests an inclusive, better solution to conduct time series forecasting for bitcoin dataset.

6.2. Discussion and Conclusion

The Bitcoin dataset consists of "Open High Low Close (OHLC)" data from January 1, 2012 to March 31, 2021. The prediction period was short due to very high volatility of bitcoin prices, i.e. from December 29, 2020 to December 31, 2020. ARIMA, XG Boost, LSTM and Facebook Prophet are the most common machine learning models used for timeseries forecasting. "Root Mean Squared Error (RMSE)" and "Mean Absolute Error (MAE)" are widely used parameters for this purpose. LSTM has been found to be ideal model among other three algorithms with RMSE of 0.3017 and MAE of 0.4414. The way LSTM performed better than other models shows that it considers different decompositions like "Sessional and Residual Trend" and "Regular "Trend". Even with higher deviation of prices than normal, LSTM is found to be very accurate for predictions for both long-term and short-term prices.

6.3. Contributions

All of the cryptocurrencies share same history and there is a lot of room for experimentation and improvement if other cryptocurrencies are added for analysis. In addition, all the models discussed here rely on previous data, i.e. past prices. So, it is not feasible to depend on these forecasts for trading in cryptocurrencies. There are so many factors influencing crypto prices. It is important to consider other factors for prediction to mitigate risks and get better results. Adding governance would help in better prediction accuracy in future advancements. For example, gathering data on

Twitter's trending page would provide data on public sentiments about investment in Bitcoin, which would also affect the demand and drive its price across the world. Similarly, analyzing Reddit posts would also help to go deeper and capture sentiments of the investors and common public. These factors must be considered for training models to further improve accuracy and efficiency.

Cryptocurrency is a digital asset derived on the protocols and technologies of blockchain which has a secure platform and relies on a decentralized network for transaction which minimizes the archives of fake processes. It can be explained further by comparing it to conventional currencies which have enough control of central authority and blockchain ecosystem has garnered enough attentions over the years for highest exchange rate. One of the major challenges in the market is prediction of finances, which have highest degree of relationships, uncertainty, and quietness.

The Long Short-Term Memory (LSTM) model proves to be the most effective when handling volatile and hard-to-predict data like Bitcoin prices. This Bitcoin dataset consisted of extremely volatile and abnormal time series data. LSTM model gives a very good RMSE, and MAE values as compared to the other models used and the models had a little difficult time predicting for 30 days of prices. In addition, could have picked better and more informed parameters for some of these models and libraries in order to make the results fairer. However, the LSTM model's low RMSE and MAE proves how powerful neural networks can be in Machine Learning.

In the basis of market innovation and position of digital assets, a lot of studies developed "exchange prediction" issue. AI algorithms like "Support Vector Regression (SVR)", "Artificial Neural Network (ANN)", and "Bayesian Neural Network (BNN)" have been applicable for prediction of bitcoin prices. There are several AI algorithms which enabled extracting new patterns and hidden data from big data with the need for any amount of information related to the dataset. The digital economics have led to a great interruption in virtually all financial economics and systems.

6.4. Future Work

Stock brokers and crypto traders can have upper hand in the market with exchange rate prediction of digital coins in the near future. The algorithm returned with precise outcomes deploying the trained model. XG Boost comes up with great results in comparison to other algorithms for predicting exchange rate of regular records of Ether, Litecoin and Bitcoin. The prediction model is evaluated on the basis of MAE, RMSE and R^2 score. Additionally, there is a smaller record of RMSE in XG Boost than others. The finding is based on prediction of exchange rate on daily basis with various parameters and resources.

In this process, "knowledge discovery" is based on incomplete cryptocurrency data in social media and crypto data is analyzed to achieve better results for more processing results. The prices of different cryptocurrencies can be forecasted on the basis of historical data with machine learning models, and it can give great results. Considering the plots and values, LSTM performed the best as compared to the other models to forecast cryptocurrency prices. However, these models are not limited to price prediction only. They can perform better for several other predictions and forecasts and they could be used in financial markets or stocks. Overall, all the models discussed here came up with solid performance.

References

Abraham, M., 2020. Studying the patterns and long-run dynamics in cryptocurrency prices. Journal of Corporate Accounting & Finance, 31(3), pp.98-113.

Abu-Mostafa, Y.S. and Atiya, A.F., 1996. Introduction to financial forecasting. Applied intelligence, 6(3), pp.205-213.

Aggarwal, P.K., Jain, P. and Vajpayee, P., 2021. Time Series Forecasting of Bitcoin Trading Price using ARIMA, FB Prophet, XG Boost and LSTM. Design Engineering, pp.11277-11299.

Aharon, D.Y. and Qadan, M., 2019. Bitcoin and the day-of-the-week effect. *Finance Research Letters*, *31*.

Alessandretti, L., ElBahrawy, A., Aiello, L.M. and Baronchelli, A., 2018. Anticipating cryptocurrency prices using machine learning. *Complexity*, *2018*.

Anghel, D.G., 2021. A reality check on trading rule performance in the cryptocurrency market: Machine learning vs. technical analysis. Finance Research Letters, 39, p.101655.

Bação, P., Duarte, A.P., Sebastião, H. and Redzepagic, S., 2018. Information transmission between cryptocurrencies: does bitcoin rule the cryptocurrency world?. *Scientific Annals of Economics and Business*, *65*(2), pp.97-117.

Bariviera, A.F., 2017. The inefficiency of Bitcoin revisited: A dynamic approach. *Economics Letters*, *161*, pp.1-4.

Baur, D.G. and Dimpfl, T., 2017. Realized bitcoin volatility. SSRN, 2949754, pp.1-26.

Baur, D.G., Hong, K. and Lee, A.D., 2018. Bitcoin: Medium of exchange or speculative assets?. *Journal of International Financial Markets, Institutions and Money*, *54*, pp.177-189.

Ben-Hur, A. and Weston, J., 2010. A user's guide to support vector machines. In *Data mining techniques for the life sciences* (pp. 223-239). Humana Press.

Borges, T.A. and Neves, R.F., 2020. Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods. Applied Soft Computing, 90, p.106187.

Bouoiyour, J. and Selmi, R., 2015. What does Bitcoin look like?. *Annals of Economics & Finance*, *16*(2).

Bouri, E., Molnár, P., Azzi, G., Roubaud, D. and Hagfors, L.I., 2017. On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier?. *Finance Research Letters*, *20*, pp.192-198.

Caporale, G.M., Plastun, A., 2019. The day of the week effect in the cryptocurrency market. *Finance Res Lett* 31:258–269.

Catania, L., Grassi, S. and Ravazzolo, F., 2019. Forecasting cryptocurrencies under model and parameter instability. *International Journal of Forecasting*, *35*(2), pp.485-501.

Chan, W.N., 2020. Time series data mining: comparative study of ARIMA and Prophet methods for forecasting closing prices of Myanmar Stock Exchange. J. Comput. Appl. Res, 1, pp.75-80.

Charfeddine, L. and Maouchi, Y., 2019. Are shocks on the returns and volatility of cryptocurrencies really persistent?. *Finance Research Letters*, *28*, pp.423-430.

Chaudhari, A. (2020). Forecasting Cryptocurrency Prices using Machine Learning (Doctoral dissertation, Dublin, National College of Ireland).

Cheah, E.T. and Fry, J., 2015. Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics letters*, *130*, pp.32-36.

Chen, C., Liu, L. and Zhao, N., 2020a. Fear sentiment, uncertainty, and bitcoin price dynamics: The case of COVID-19. *Emerging Markets Finance and Trade*, *56*(10), pp.2298-2309.

Chen, C.Y.H. and Hafner, C.M., 2019. Sentiment-induced bubbles in the cryptocurrency market. Journal of Risk and Financial Management, 12(2), p.53.

Chen, Z., Li, C. and Sun, W., 2020b. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, p.112395.

Cheung, A., Roca, E. and Su, J.J., 2015. Crypto-currency bubbles: an application of the Phillips–Shi–Yu (2013) methodology on Mt. Gox bitcoin prices. *Applied Economics*, *47*(23), pp.2348-2358.

Chivukula, R., & Lakshmi, T. J. (2020). Cryptocurrency Price Prediction: A Machine Learning Approach. Sensors & Transducers, 244(5), 44-47.

Chowdhury, R., Rahman, M.A., Rahman, M.S. and Mahdy, M.R.C., 2020. An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. Physica A: Statistical Mechanics and its Applications, 551, p.124569.

Ciaian, P., Rajcaniova, M. and Kancs, D.A., 2016. The economics of BitCoin price formation. *Applied economics*, *48*(19), pp.1799-1815.

Corbet S, Lucey B, Yarovaya L (2018a) Datestamping the Bitcoin and Ethereum bubbles. Finance Res Lett 26:81–88

Corbet, S., Meegan, A., Larkin, C., Lucey, B. and Yarovaya, L., 2018b. Exploring the dynamic relationships between cryptocurrencies and other financial assets. *Economics Letters*, *165*, pp.28-34.

Dastgir, S., Demir, E., Downing, G., Gozgor, G. and Lau, C.K.M., 2019. The causal relationship between Bitcoin attention and Bitcoin returns: Evidence from the Copula-based Granger causality test. *Finance Research Letters*, *28*, pp.160-164.

Derbentsev, V., Datsenko, N., Stepanenko, O. and Bezkorovainyi, V., 2019. Forecasting cryptocurrency prices time series using machine learning approach. In *SHS Web of Conferences* (Vol. 65, p. 02001). EDP Sciences.

Derbentsev, V., Datsenko, N., Stepanenko, O. and Bezkorovainyi, V., 2019. Forecasting cryptocurrency prices time series using machine learning approach. In SHS Web of Conferences (Vol. 65, p. 02001). EDP Sciences.

Dorfleitner, G. and Lung, C., 2018. Cryptocurrencies from the perspective of euro investors: a re-examination of diversification benefits and a new day-of-the-week effect. *Journal of Asset Management*, *19*(7), pp.472-494.

Dwyer, G.P., 2015. The economics of Bitcoin and similar private digital currencies. *Journal of financial stability*, *17*, pp.81-91.

Fang, F., Ventre, C., Basios, M., Kanthan, L., Martinez-Rego, D., Wu, F. and Li, L., 2022. Cryptocurrency trading: a comprehensive survey. *Financial Innovation*, *8*(1), pp.1-59.

Fang, T., Su, Z. and Yin, L., 2020. Economic fundamentals or investor perceptions? The role of uncertainty in predicting long-term cryptocurrency volatility. International Review of Financial Analysis, 71, p.101566.

Fang, W.X., Lan, P.C., Lin, W.R., Chang, H.C., Chang, H.Y. and Wang, Y.H., 2019, December. Combine facebook prophet and LSTM with BPNN forecasting financial markets: the morgan Taiwan index. In 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) (pp. 1-2). IEEE.

Foley, S., Karlsen, J.R. and Putniņš, T.J., 2019. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?. *The Review of Financial Studies*, *32*(5), pp.1798-1853.

Foley, S., Karlsen, J.R. and Putniņš, T.J., 2019. Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies?. *The Review of Financial Studies*, *32*(5), pp.1798-1853.

Gibert Llauradó, D., Mateu Piñol, C. and Planes Cid, J., 2020. The rise of machine learning for detection and classification of malware: Research developments, trends and challenge. Journal of Network and Computer Applications, 2020, vol. 153, 102526.

Gkillas, K. and Katsiampa, P., 2018. An application of extreme value theory to cryptocurrencies. *Economics Letters*, *164*, pp.109-111.

Gurdgiev, C. and O'Loughlin, D., 2020. Herding and anchoring in cryptocurrency markets: Investor reaction to fear and uncertainty. Journal of Behavioral and Experimental Finance, 25, p.100271.

Gurdgiev, C. and O'Loughlin, D., 2020. Herding and anchoring in cryptocurrency markets: Investor reaction to fear and uncertainty. *Journal of Behavioral and Experimental Finance*, *25*, p.100271.

Hyun, S., Lee, J., Kim, J.M. and Jun, C., 2019. What coins lead in the cryptocurrency market: using Copula and neural networks models. *Journal of Risk and Financial Management*, *12*(3), p.132.

Iqbal, M., Iqbal, M.S., Jaskani, F.H., Iqbal, K. and Hassan, A., 2021. Time-series prediction of cryptocurrency market using machine learning techniques. EAI Endorsed Transactions on Creative Technologies, p.e4.

Karakoyun, E.S. and Cibikdiken, A.O., 2018, May. Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting. In *The 13th multidisciplinary academic conference in Prague* (Vol. 2018, pp. 171-180).

Kavitha, H., Sinha, U.K. and Jain, S.S., 2020, January. Performance Evaluation of Machine Learning Algorithms for Bitcoin Price Prediction. In 2020 Fourth International Conference on Inventive Systems and Control (ICISC) (pp. 110-114). IEEE.

Kim, Y.B., Kim, J.G., Kim, W., Im, J.H., Kim, T.H., Kang, S.J. and Kim, C.H., 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, *11*(8), p.e0161197.

Koker, T.E. and Koutmos, D., 2020. Cryptocurrency trading using machine learning: A Technical note.

Kou, G., Lu, Y., Peng, Y. and Shi, Y., 2012. Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, *11*(01), pp.197-225.

Kou, G., Peng, Y. and Wang, G., 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information sciences*, *275*, pp.1-12.

Kristjanpoller, W. and Minutolo, M.C., 2018. A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. *Expert Systems with Applications*, *109*, pp.1-11.

Kristoufek, L., 2013. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific reports*, *3*(1), pp.1-7.

Kristoufek, L., 2015. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PloS one*, *10*(4), p.e0123923.

Kumar, A.P. and Sunny, S., 2021. Comparative Analysis Of Machine Learning Models For Predicting Bitcoin Price Rate. International Research Journal of Modernization in Engineering Technology and Science, 03(09).

Kwon, D.H., Kim, J.B., Heo, J.S., Kim, C.M. and Han, Y.H., 2019. Time series classification of cryptocurrency price trend based on a recurrent LSTM neural network. Journal of Information Processing Systems, 15(3), pp.694-706.

Lahmiri, S. and Bekiros, S., 2019. Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, *118*, pp.35-40.

Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos, Solitons & Fractals, 118, 35-40.

Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos, Solitons & Fractals, 118, 35-40.

Li, T., Kou, G. and Peng, Y., 2020. Improving malicious URLs detection via feature engineering: Linear and nonlinear space transformation methods. *Information Systems*, *91*, p.101494.

Li, T., Kou, G., Peng, Y. and Shi, Y., 2017. Classifying with adaptive hyper-spheres: An incremental classifier based on competitive learning. *IEEE transactions on systems, man, and cybernetics: systems*, *50*(4), pp.1218-1229.

Li, T.R., Chamrajnagar, A.S., Fong, X.R., Rizik, N.R. and Fu, F., 2019. Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. Frontiers in Physics, 7, p.98.

Li, X. and Wang, C.A., 2017. The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. *Decision support systems*, *95*, pp.49-60.

Li, X. and Wang, C.A., 2017. The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. *Decision support systems*, *95*, pp.49-60.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R news*, *2*(3), pp.18-22.

Liew, J., Li, R.Z., Budavári, T. and Sharma, A., 2019. Cryptocurrency investing examined. *The Journal of The British Blockchain Association*, p.8720.

Livieris, I.E., Pintelas, E., Stavroyiannis, S. and Pintelas, P., 2020. Ensemble deep learning models for forecasting cryptocurrency time-series. Algorithms, 13(5), p.121.

Mallqui, D.C. and Fernandes, R.A., 2019. Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. Applied Soft Computing, 75, pp.596-606.

Mallqui, D.C. and Fernandes, R.A., 2019. Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, *75*, pp.596-606.

Marsh, A. 2018. *'What is Bitcoin?' topped Google's 2018 trending list*, *mint*. Available at: https://www.livemint.com/Money/O8Jamo3u4HyqjoWxY0wBoN/What-is-Bitcoin-topped-Googles-2018-trending-list.html (Accessed: 30 May 2022).

Mehtab, S. and Sen, J., 2020. A time series analysis-based stock price prediction using machine learning and deep learning models. International Journal of Business Forecasting and Marketing Intelligence, 6(4), pp.272-335.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C. and Lin, C., 2015. Misc functions of the department of statistics, probability theory group (formerly: E1071). *Package e1071. TU Wien*.

Mikhaylov, A., 2020. Cryptocurrency market analysis from the open innovation perspective. Journal of Open Innovation: Technology, Market, and Complexity, 6(4), p.197.

Mittal, A., Dhiman, V., Singh, A. and Prakash, C., 2019, August. Short-term bitcoin price fluctuation prediction using social media and web search data. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-6). IEEE.

Mohapatra, S., Ahmed, N. and Alencar, P., 2019, December. Kryptooracle: A real-time cryptocurrency price prediction platform using twitter sentiments. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 5544-5551). IEEE.

Mudassir, M., Bennbaia, S., Unal, D. and Hammoudeh, M., 2020. Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. Neural computing and applications, pp.1-15.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system. Decentralized Business Review, p.21260.

Omane-Adjepong, M. and Alagidede, I.P., 2019. Multiresolution analysis and spillovers of major cryptocurrency markets. *Research in International Business and Finance*, *49*, pp.191-206.

Panagiotidis, T., Stengos, T. and Vravosinos, O., 2018. On the determinants of bitcoin returns: A LASSO approach. *Finance Research Letters*, *27*, pp.235-240.

Panagiotidis, T., Stengos, T. and Vravosinos, O., 2019. The effects of markets, uncertainty and search intensity on bitcoin returns. *International Review of Financial Analysis*, *63*, pp.220-242.

Pant, D.R., Neupane, P., Poudel, A., Pokhrel, A.K. and Lama, B.K., 2018, October. Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)* (pp. 128-132). IEEE.

Parizi, R.M., Dehghantanha, A., Choo, K.K.R. and Singh, A., 2018. Empirical vulnerability analysis of automated smart contracts security testing on blockchains. arXiv preprint arXiv:1809.02702.

Parkinson, M., 1980. The extreme value method for estimating the variance of the rate of return. *Journal of business*, pp.61-65.

Patel, J., Shah, S., Thakkar, P. and Kotecha, K., 2015. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert systems with applications*, *42*(1), pp.259-268.

Peng, Y., Albuquerque, P.H.M., de Sá, J.M.C., Padula, A.J.A. and Montenegro, M.R., 2018. The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression. *Expert Systems with Applications*, *97*, pp.177-192.

Phillips, R.C. and Gorse, D., 2017, November. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *2017 IEEE symposium series on computational intelligence (SSCI)* (pp. 1-7). IEEE.

Phillips, R.C. and Gorse, D., 2018. Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PloS one*, *13*(4), p.e0195200.

Polasik, M., Piotrowska, A.I., Wisniewski, T.P., Kotkowski, R. and Lightfoot, G., 2015. Price fluctuations and the use of bitcoin: An empirical inquiry. *International Journal of Electronic Commerce*, *20*(1), pp.9-49.

Poloniex (2019). https://www.poloniex.com/. [Online].

Poongodi, M., Sharma, A., Vijayakumar, V., Bhardwaj, V., Sharma, A.P., Iqbal, R. and Kumar, R., 2020. Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. Computers & Electrical Engineering, 81, p.106527.

Poongodi, M., Sharma, A., Vijayakumar, V., Bhardwaj, V., Sharma, A.P., Iqbal, R. and Kumar, R., 2020. Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. Computers & Electrical Engineering, 81, p.106527.

Pyo, S. and Lee, J., 2020. Do FOMC and macroeconomic announcements affect Bitcoin prices?. *Finance Research Letters*, *37*, p.101386.

Quandl (2019). https://www.quandl.com/. [Online].

Rathan, K., Sai, S.V. and Manikanta, T.S., 2019, April. Crypto-currency price prediction using decision tree and regression techniques. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 190-194). IEEE.

Sattar, A., Ertuğrul, Ö.F., Gharabaghi, B., McBean, E.A. and Cao, J., 2019. Extreme learning machine model for water network management. Neural Computing and Applications, 31(1), pp.157-169.

Saxena, A., Sukumar, T.R., 2018. Predicting bitcoin price using LSTM and compare its predictability with ARIMA model. *International Journal of Pure and Applied Mathematics*, *119*(17), pp.2591-2600.

Schmidhuber, J. and Hochreiter, S., 1997. Long short-term memory. *Neural Comput*, *9*(8), pp.1735-1780.

Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K. and Soman, K.P., 2017, September. Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) (pp. 1643-1647). IEEE.

Sen, J. and Chaudhuri, T.D., 2018, December. Stock price prediction using machine learning and deep learning frameworks. In Proceedings of the 6th International Conference on Business Analytics and Intelligence, Bangalore, India (pp. 20-22).

Shah, D., Isah, H. and Zulkernine, F., 2019. Stock market analysis: A review and taxonomy of prediction techniques. *International Journal of Financial Studies*, *7*(2), p.26.

Singh, H. and Agarwal, P., 2018, August. Empirical analysis of bitcoin market volatility using supervised learning approach. In *2018 Eleventh International Conference on Contemporary Computing (IC3)* (pp. 1-5). IEEE.

Singh, S., Madan, T.K., Kumar, J. and Singh, A.K., 2019, July. Stock market forecasting using machine learning: Today and tomorrow. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) (Vol. 1, pp. 738-745). IEEE.

Smola, A.J. and Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and computing*, *14*(3), pp.199-222.

Sovbetov, Y., 2018. Factors influencing cryptocurrency prices: Evidence from bitcoin, ethereum, dash, litcoin, and monero. *Journal of Economics and Financial Analysis*, *2*(2), pp.1-27.

Stavroyiannis, S. and Babalos, V., 2019. Herding behavior in cryptocurrencies revisited: Novel evidence from a TVP model. *Journal of Behavioral and Experimental Finance*, *22*, pp.57-63.

Steinert, L. and Herff, C., 2018. Predicting altcoin returns using social media. *PloS one*, *13*(12), p.e0208119.

Sun, J., Zhou, Y. and Lin, J., 2019, May. Using machine learning for cryptocurrency trading. In 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS) (pp. 647-652). IEEE.

Sun, X., Liu, M., & Sima, Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. Finance Research Letters, 32, 101084.

Tay, F.E. and Cao, L., 2001. Application of support vector machines in financial time series forecasting. *omega*, *29*(4), pp.309-317.

Taylor, S.J. and Letham, B., 2018. Forecasting at scale. *The American Statistician*, *72*(1), pp.37-45.

Torgo, L., 2016. Data mining with R: Learning with case studies, Data Mining with R: Learning with Case Studies.

Urquhart, A., 2016. The inefficiency of Bitcoin. *Economics Letters*, *148*, pp.80-82.

Wadalkar, Y., Sai Tarun, Y., Singhal, J. and Sonkusare, P., 2021. A Comparative Analysis based approach for Bitcoin Price Forecasting. International Journal of Engineering Research & Technology (IJERT), 10(4).

Wang, L., Zou, H., Su, J., Li, L. and Chaudhry, S., 2013. An ARIMA-ANN hybrid model for time series forecasting. Systems Research and Behavioral Science, 30(3), pp.244-259.

Wen, F., Xu, L., Ouyang, G. and Kou, G., 2019. Retail investor attention and stock price crash risk: evidence from China. *International Review of Financial Analysis*, *65*, p.101376.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

Wimalagunaratne, M. and Poravi, G., 2018, May. A predictive model for the global cryptocurrency market: a holistic approach to predicting cryptocurrency prices. In *2018 8th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)* (pp. 78-83). IEEE.

Wu, C.H., Lu, C.C., Ma, Y.F. and Lu, R.S., 2018, November. A new forecasting framework for bitcoin price with LSTM. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 168-175). IEEE.

Yaga, D., Mell, P., Roby, N. and Scarfone, K., 2019. Blockchain technology overview. *arXiv preprint arXiv:1906.11078*.

Yenidoğan, I., Çayir, A., Kozan, O., Dağ, T. and Arslan, Ç., 2018, September. Bitcoin forecasting using ARIMA and PROPHET. In *2018 3rd international conference on computer science and engineering (UBMK)* (pp. 621-624). IEEE.

Yermack, D., 2015. Is Bitcoin a real currency? An economic appraisal. In *Handbook of digital currency* (pp. 31-43). Academic Press.

Yiying, W. and Yeze, Z., 2019, March. Cryptocurrency price analysis with artificial intelligence. In 2019 5th International Conference on Information Management (ICIM) (pp. 97-101). IEEE.

Yiying, W. and Yeze, Z., 2019, March. Cryptocurrency price analysis with artificial intelligence. In 2019 5th International Conference on Information Management (ICIM) (pp. 97-101). IEEE.

Yu, H. and Kim, S., 2012. SVM Tutorial-Classification, Regression and Ranking. *Handbook of Natural computing*, *1*, pp.479-506.

Yu, J.H., Kang, J. and Park, S., 2019. Information availability and return volatility in the bitcoin market: analyzing differences of user opinion and interest. *Information Processing & Management*, *56*(3), pp.721-732.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, pp.159-175.

Zhu, Y., Dickinson, D. and Li, J., 2017. Analysis on the influence factors of Bitcoin's price based on VEC model. *Financial Innovation*, *3*(1), pp.1-13.

Hu, Z., Zhao, Y. and Khushi, M., 2021. A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, *4*(1), p.9.

Poongodi, M., Sharma, A., Vijayakumar, V., Bhardwaj, V., Sharma, A.P., Iqbal, R. and Kumar, R., 2020. Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system. *Computers & Electrical Engineering*, *81*, p.106527.

Iqbal, M., Iqbal, M.S., Jaskani, F.H., Iqbal, K. and Hassan, A., 2021. Time-series prediction of cryptocurrency market using machine learning techniques. *EAI Endorsed Transactions on Creative Technologies*, p.e4.

Mallqui, D.C. and Fernandes, R.A., 2019. Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques. *Applied Soft Computing*, *75*, pp.596-606.

Sebastião, H. and Godinho, P., 2021. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financial Innovation*, *7*(1), pp.1-30.

Derbentsev, V., Datsenko, N., Stepanenko, O. and Bezkorovainyi, V., 2019. Forecasting cryptocurrency prices time series using machine learning approach. In *SHS Web of Conferences* (Vol. 65, p. 02001). EDP Sciences.

Yiying, W. and Yeze, Z., 2019, March. Cryptocurrency price analysis with artificial intelligence. In 2019 5th International Conference on Information Management (ICIM) (pp. 97-101). IEEE.

Lahmiri, S. and Bekiros, S., 2019. Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, *118*, pp.35-40.

Rathan, K., Sai, S.V. and Manikanta, T.S., 2019, April. Crypto-currency price prediction using decision tree and regression techniques. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 190-194). IEEE.

Borges, T.A. and Neves, R.F., 2020. Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods. *Applied Soft Computing*, *90*, p.106187.

Kwon, D.H., Kim, J.B., Heo, J.S., Kim, C.M. and Han, Y.H., 2019. Time series classification of cryptocurrency price trend based on a recurrent LSTM neural network. *Journal of Information Processing Systems*, *15*(3), pp.694-706.

Sen, J., Mehtab, S. and Nath, G., 2020. Stock price prediction using deep learning models. *Lattice-The Machine Learning Journal*, *1*(3), pp.34-40.

Sen, J. and Chaudhuri, T.D., 2018, December. Stock price prediction using machine learning and deep learning frameworks. In *Proceedings of the 6th International Conference on Business Analytics and Intelligence, Bangalore, India* (pp. 20-22).

Patel, M.M., Tanwar, S., Gupta, R. and Kumar, N., 2020. A deep learning-based cryptocurrency price prediction scheme for financial institutions. *Journal of information security and applications*, *55*, p.102583.

Chen, Z., Li, C. and Sun, W., 2020. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, p.112395.

Dutta, A., Kumar, S. and Basu, M., 2020. A gated recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, *13*(2), p.23.

APPENDIX A: RESEARCH PROPOSAL

CRYPTOCURRENCY TIME SERIES FORECASTING

AVLEEN SINGH

M.SC DATA SCIENCE

Research Proposal

APRIL 2022

Abstract

Cryptocurrency is an online mechanism for payments which works independently from banks or any centralized authority. This P2P system allows users to conduct transactions on the go. There is a public ledger to record all the digital transactions about cryptocurrency funds, which is kept in digital wallets. These are basically digital assets like securities, stocks, gold, etc. Market sentiment is the major factor which determines their value, which is simply known as demand and its availability or supply. From being known completely to experts and insiders to mainstream media attention these days, crypto assets like bitcoin, Litecoin, Ether, have come a long way. These assets can be seen only digitally, and only digital addresses are needed to transfer them. Unlike traditional currencies which are known to common public and rely on private bank or financial institutions for transactions and may be kept in physical form, cryptocurrency has no physical form or central authority to control. This study is aimed to compare and forecast cryptocurrency price with various machine learning models like FBProphet, XG Boost, LSTM and ARIMA and come up with the ideal model for forecasting bitcoin prices. This thesis is a milestone for those who are planning to invest in cryptocurrency, and it will also help experts to predict crypto prices early on.

Keywords – "*LSTM, ARIMA, Cryptocurrency, FBProphet, XG Boost, time series forecasting, machine learning models*"

Table of Contents

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADF	Augmented Dickey-Fuller
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
BTC	Bitcoin
EDA	Exploratory Data Analysis
ICO	Initial Coin Offer
IDE	Integrated Development Environment
KDE	Kernel Density Estimate
KNN	K-Nearest Neighbor
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LSTM	Long Short-term Memory
OHLC	Open, High, Low, Close
OS	Operating System
RAM	Random Access Memory
RNN	Recurrent Neural Network
SHA-256	Secure Hash Algorithm 256-bit
USD	United States Dollar

1. Background

Cryptocurrency is a type of digital currency that serves as a means of exchange for ensuring transaction security using encryption. In 2009, Bitcoin, the most widely used cryptocurrency, was the first decentralized currency. Since then, a number of other cryptocurrencies have emerged. Although it has just been nine years since its inception, the cryptocurrency market has grown fast in recent years, particularly since 2017. Bitcoin had a unit price of less than $1,000 at the start of 2017. By the end of the year, the unit price had risen to more than $13,000, and the market value had climbed by more than 12 times. At the end of December 2017, there were more than thirty cryptocurrencies with a market value of more than $1 billion. Despite the fact that the cryptocurrency market crashed in 2018, there are still thirteen cryptocurrencies with a market capitalization of more than $1 billion. Cryptocurrency has grown into a significant part of the global financial sector that cannot be ignored any longer. As a growing alternative asset, cryptocurrency can be invested in a variety of ways. (Using machine learning for cryptocurrency trading, 2019).

Bitcoin is the longest running and most well-known cryptocurrency, first released as open source in 2009 by the anonymous Satoshi Nakamoto. Bitcoin was created to be an electronic peer-to-peer cash system, but has also attracted crypto-curious investors as a store-of-value currency, comparable to gold. Bitcoin serves as a decentralized medium of digital exchange, with transactions verified and recorded in a public distributed ledger (the block chain) without the need for a trusted record keeping authority or central intermediary. Transaction blocks contain a SHA-256 cryptographic hash of previous transaction blocks, and are thus "chained" together, serving as an immutable record of all transactions that have ever occurred. As with any currency/commodity on the market, bitcoin trading and financial instruments soon followed public adoption of bitcoin and continue to grow. Bitcoins are created as a reward for a process known as mining. They can be exchanged for other currencies, products, and services. Bitcoin has been criticized for its use in illegal transactions, the large amount of electricity used by mining, price volatility, and thefts from exchanges. In September 2021, El Salvador officially adopted Bitcoin as legal tender, in the face of internal and international criticism, becoming the first nation to do so.

During the last decade, the number of Cryptocurrencies bought and exchanged has increased tremendously, with a daily market valuation of hundreds of billions of dollars (1 trillion to January 2021). Distributed ledgers and block chains are at the heart of these crypto markets' technology, and cryptocurrencies rely on them to be mined, i.e. produced, and exchanged among users and other stakeholders. These technologies ensure anonymity and secure value transfer, resulting in the emergence of the aforementioned crypto marketplaces. The number of live projects connected to block chain technologies and distributed ledgers surpassed 7000 in January 2021, and the trend is continuing to grow. Bitcoin is the first and most well-known cryptocurrency, and its associated tokens dominate the crypto-markets in terms of both market volume and valuation. Due to the extreme volatility of crypto-markets, investors and stakeholders have been drawn to this technology, its prospects, the potential for constant value increase, and the excess returns on their investments since the inception of Bitcoin in 2009. Due to the novelty of the platforms, millions of users, and investors in the area, valuing and pricing cryptocurrencies and digitally native tokens and sheets is a difficult task. We concentrated our investigation on Bitcoin for the reasons stated above, as well as the fact that it has proven to be both safe and resilient to network attacks. Several researchers have focused their efforts on quantitative investigations to extract information from cryptocurrency price time series and forecasting price insights or to predict the next most likely jump in value, ranging from theoretical models of digital token pricing and adoption to machine learning and neural network-driven price and return predictions. The current state of study in this sector has provided insights into the maturity, efficiency, and structure of cryptocurrency marketplaces. The investigation of reasons underlying the extreme volatility of cryptocurrencies has received special attention, ranging from extrapolating the mechanisms causing the swings to model estimation: Some study has found a strong link between global economic activity and trade volume, as well as open source development communities that promote cryptocurrencies (Investigation of Block chain Cryptocurrencies Price Movements through Deep Learning a Comparative Analysis, 2021).

2. Problem Statement

Currently, research on cryptocurrency price forecasts is at an early stage. There are several related articles that study cryptocurrency transactions from different methods and perspectives.

Some research that predicts the price of cryptocurrencies through machine learning techniques and some people have analyzed the factors affecting the popularity of cryptocurrency. Some people have built different models to describe the trend of cryptocurrency (Using machine learning for cryptocurrency trading, 2019).

Cryptocurrencies have recently experienced a surge in popularity and price. Cryptocurrency prices, as well as various internet and social media characteristics, have been found to be linked to time periods. Furthermore, cryptocurrencies have been found to be subject to bubble-like price increase phases. However, the notion studied is that parameters like price and online variables are influenced by the consumer regime. Based on 1-min price changes reported since 2012, the fluctuating features of the fast evolving Bitcoin market are measured over selected sub-periods in terms of return distributions, volatility autocorrelation, Hurst exponents, and multi-scaling effects. These changes were investigated in relation to the phenomenon of Initial Coin Offerings (ICOs), which boosted the cryptocurrency market environment. Despite the fact that high-frequency price data has been available from the beginning of trading, Bitcoin offers a unique look into the statistical characteristics of market maturation. Money's invention was one of civilization's most significant achievements in history, and it simultaneously conditioned their continued progress. Quantitative studies aimed at extracting information from cryptocurrency price time series and forecasting price drivers or the next most likely jump span from theoretical models of digital token pricing and adoption to machine learning and neural network-driven price and return projections. Other studies focus into the relationships between cryptocurrency development communities and price movements, as well as the sociological variables that promote the development of cryptocurrencies. The maturity, efficiency, and structure of the bitcoin markets have all been studied (Investigation of Blockchain Cryptocurrencies Price Movements Through Deep Learning A Comparative Analysis, 2021).

In the financial and medical fields, the Artificial Neural Network (ANN) is one of the most powerful predicting systems. Pattern recognition, prediction, and classification are some of the tasks that an artificial neural network can accomplish with high accuracy. A powerful universal approximator, the artificial neural network can estimate any continuous function. These continuous functions can be approximated to any desired level of accuracy, resulting in a very accurate prediction. Because of the reasons outlined above, artificial neural networks are a better alternative to traditional statistical models. A literature evaluation of comparison studies between ANN and traditional statistical models indicated that traditional approaches outperformed ANN in just 18% of the cases, whereas ANN models performed well or outperformed in 72 percent of the cases. Due to the highly non-linear and volatile nature of the financial market, it is unsurprising that neural networks have been garnering more attention in the field of finance, particularly for time series forecasting. Because Bitcoin is a relatively new technology, there are only a few price prediction models available. It works with time series data and has developed three data sets: 120, 60, and 30 minutes. Then they use a random forest classifier to create three linear models from their datasets. These three models are linearly merged to estimate the price of Bitcoin. The non-linear autoregressive exogenous and multi-layer perceptron are compared.

They used the MATLAB neural network toolkit to assess and develop the network. Despite the fact that non-linear autoregressive exogenous outperforms the multi-layer perceptron, the authors believe that the multi-layer perceptron can be utilised for stock market prediction because it can be readily adapted for different Bitcoin parameters. When compared to a standard artificial neural network, the author's supported vector machine and artificial neural network produced 55 percent better outcomes. Multi-layer perceptron has been used in several research, although it only analyses one observation at a time. Long short-term memory is an improved version of recurrent neural networks. In addition to having a large short-term memory, you may pick which data to forget and which data to remember based on the importance and weight of each feature. The author has created a Long Short-Term Memory using Recurrent Neural Network, which is explained in this paper and used for implementation. Because the Long Short-Term Memory includes a feedback network in addition to the feedforward neural network, it can be used for general-purpose computation. One small disadvantage of extended short-term memory is that it necessitates a lot of calculation, which can make processing slow at times.

On a large-scale dataset, the author constructed a neural network and compared autoregressive integrated moving average, recurrent neural network, and zhang's hybrid models, concluding that zhang's model provided the best prediction results. Based on the method of modelling and prediction of Bitcoin values, the author conducted an empirical study that compares the Bayesian

Neural Network to other linear and non-linear benchmark models. The authors employed a Genetic algorithm based selective neural network to investigate the association between Bitcoin variables such as transaction volume and cost per transaction and the change in Bitcoin price the next day using a generalised linear model and Bayesian regression (Performance Evaluation of Machine Learning Algorithms for Bitcoin Price Prediction, 2020). The cryptocurrency market has seen its rise and fall in the past few years. With a variety of coins being exchanged for real money, it is important to know the trend in the coin price. In this research, will compare the different models (like LSTM, XG Boost, Facebook Prophet and ARIMA) and provide the best suited model to predict or forecast the prices of Bitcoin.

3. Aim and Objectives

The aim of Crypto Economy Organization is to unite all stakeholders of the crypto ecosystem to create and support friendly environment to develop the crypto economy in close collaboration with other existing economic ecosystems (Price Prediction and Notification System for cryptocurrency Share Market Trading, 2021).

Considering the above statement, here are the objectives of this research –

- To develop an ideal business environment and socialize with like-minded participants of crypto ecosystem actively.
- To work with policymakers and regulatory bodies to attain competitive and effective regulations for “crypto ecosystem” and promote the usage, integration, and development of various technologies of distributed ledger and blockchain.
- To recommend participants of “crypto ecosystem” the activities related to law, business, education, and finance to build mutual trust and harmony among participants and global and regulatory bodies.
- To use “time series forecasting” with methods for analyzing “time series information” to gather eloquent data and other features.

4. Significance of Study

Bitcoin is a type of digital or virtual currency. Block chain, like other cryptocurrencies, is the foundation of Bitcoin. That is to say, it is completely decentralized. Bitcoin cannot be controlled or regulated by any central bank, government, or authority. The Bitcoin network is maintained by a peer-to-peer community of networked computers. Each user is a node in the network, and they all have the same rights. Bitcoin is amongst the most reliable and safer digital currency currently available. We just ought to trade in the best possible forms in an environment where there are too many conmen and looters. Cryptocurrencies provide us with certainty, rendering them a desirable investment opportunity both now and in the future. The strategies of bitcoins are also another justification for their rapid acceptance and the importance of cryptocurrency. You shouldn't need to communicate with a second team when it comes to cryptocurrencies. This brings people a sense of comfort and reassurance. Since cryptocurrencies are digital currencies, they do not involve a third party. You will conduct business regardless of your venue. Block chain can be used to establish a useful digital database or a distributed public ledger. And this ledger is encrypted, and it keeps track of all Bitcoin transactions. Bitcoin mining is a process in which miners painstakingly verify each transaction. As a result, Bitcoin transactions are safe. Bitcoin has also shown to be a useful asset for storing value. It's for this reason that some people compare it to gold. Bitcoin has demonstrated its ability to resist adversities better than fiat money during economic downturns. As a result, in the event of a fiscal catastrophe, you can employ this asset type to store wealth "(Price Prediction and Notification System for cryptocurrency Share Market Trading, 2021)".

5. Scope of the Study

This study would be able to understand different prices of Bitcoin, if Bitcoin is a good investment and so on. It also helps in comparing different models and providing the best prices of Bitcoin as well as the best suitable model for prediction. Time series data is a type of experimental data that has been observed over a period of time (usually evenly spaced, like once a day or once an hour or once a minute). The data on airline ticket sales per day, for example, is a time series. However, just because a set of occurrences contains a time component does not make it a time series; for example, the dates of major aircraft tragedies are not time series because they are arbitrarily spaced. The term "point process" refers to these types of random processes. Trend, seasonality, and noise are all important characteristics of time series. Forecasting is the practice of creating future predictions based on historical and current facts (Performance Evaluation of Machine Learning Algorithms for Bitcoin Price Prediction, 2020).

As the dataset is big, so will be dividing the kernel into two parts: -

- Data Exploration
- Time Series Analysis

And for further Time Series Forecasting: -

a) Time Series Forecasting with LSTM (Long Short-term Memory).

b) Time Series Forecasting with XG Boost.

c) Time Series Forecasting with Facebook Prophet.

d) Time Series Forecasting with ARIMA (Autoregressive Integrated Moving Average).

6. Research Methodology

6.1 Introduction

This section consists of important processes like data selection, its pre-processing, transformation of the same into a logical, organized format, using various techniques, balancing dataset, and determining the performance of ML with evaluation metrics.

6.2 Dataset Description

Kaggle provided the historical Bitcoin data utilized in this investigation (Bitcoin Historical Data: Bitcoin Data at 1-Min Intervals from Select Exchanges, 2012). It contains statistics from January 2012 through March 2021, including minute-by-minute updates on the opening, closing, lowest, highest, traded volume, and weighted bitcoin price. There are 4857377 records of bitcoin movement in this dataset, with 8 columns.

The dataset is comprised of various columns and below is the description: -

- The “Close” and “Open” columns are the sign of closing and opening price of the day.

- Similarly, the lowest and highest prices of the day are mentioned in “Low” and “High” columns, respectively.
- The overall traded volume on the day is mentioned in “Volume” column.
- The “Weighted price” refers to the “trading benchmark” for traders to find out the “weighted price” for trading a security in a day based on both price and volume. It gives data to the traders on both value and trend of security.

6.3 Data Pre-Processing

Some of the “data transformation” and “data pre-processing” techniques used in classification include “Outlier Treatment”, “Class Imbalance”, “Binning Continuous Variables”, “Skewness Correlation”, “Boolean Variable Mapping”, “Duplicates Removal”, “Data Format Creation”, “Dummy Variable Creation”, “Feature Scaling”, “Data Normalization”, “Test-Train Split”, among others. Some of the “data transformation” and “data pre-processing” techniques used for “time series forecasting” are “Stationary Tests” like “ADF” and “KPSS”, “Missing Value Treatment” with “Last Observation Carried Forward”, “Mean Value Imputation” or “Next Observation Carried Forward”, “Differencing”, “Linear Interpolation”, “Test-Train Split” and “Time Series Decomposition” etc.

6.4 Flow Diagram

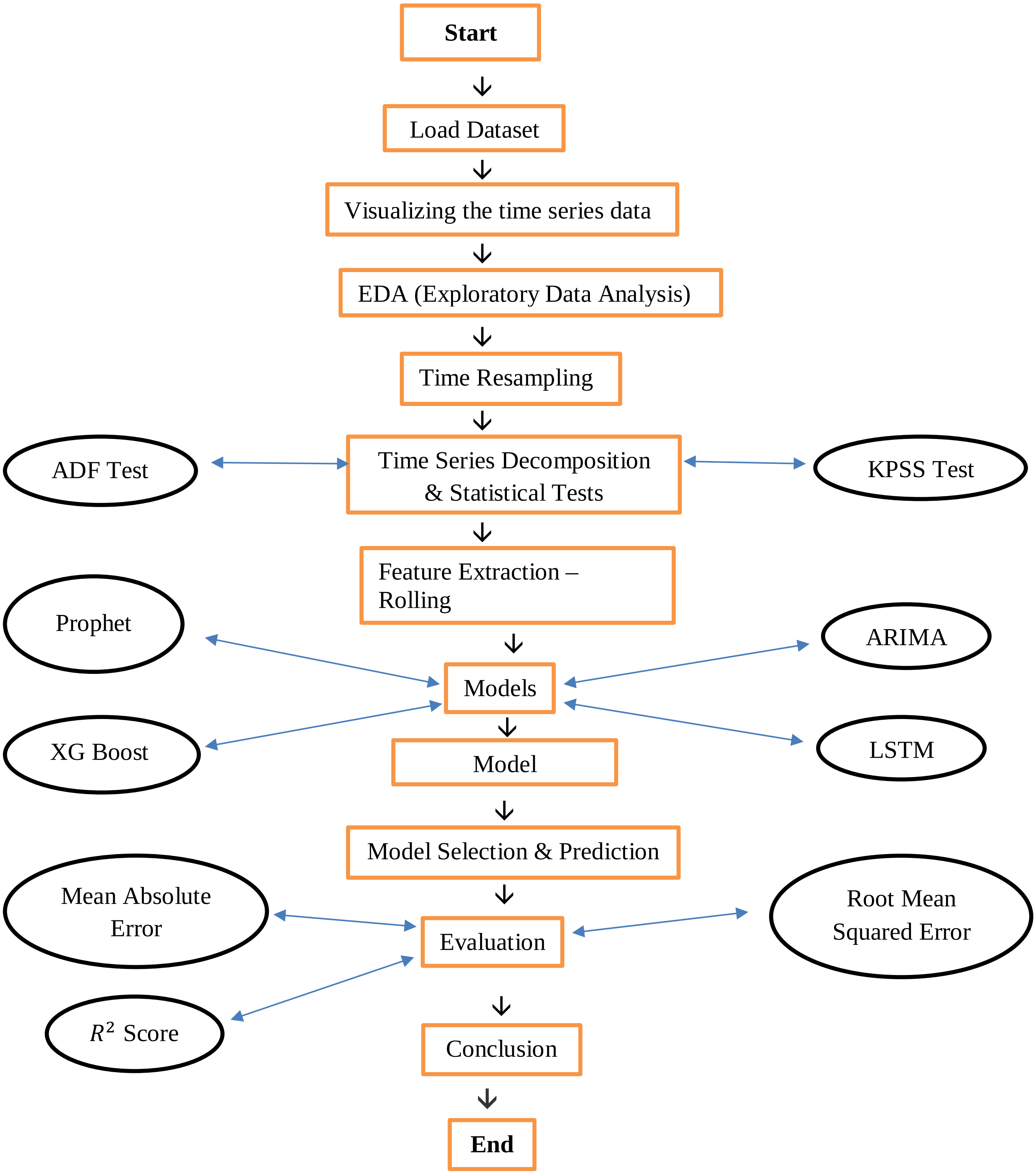


Figure 1: Flow Diagram of Research Methodology

The overall flowchart of the research is illustrated in Figure 1. In this research, the first job is to search for a proper dataset which was found from Kaggle. The next step is to load the dataset and having a first look at the dataset. Then, can check the datatypes of each columns using the info() function, post that can check the first five rows by using the head() function and then the number of rows and columns in the complete dataset by using the shape() function. Post these basic steps, can visualize the dataset by plotting them on different graphs among different columns in the dataset. By plotting different graphs, we can conclude if there are any absent values in the "time series data". If there are any "missing values" in the "time series data", then they will need to be handled. Then for non-series data, would use the imputation technique to handle them. Next, Exploratory data analysis (EDA) will be performed as the next step and in this, will be visualizing the weighted prices of Bitcoin.

Next will be visualizing the prices using "KDE (Kernel Density Estimate)". KDE is used for determining the "probability density" of a "continuous variable". It refers to the "probability density" at diverse values in a consistent variable. Once visualizing using KDEs is done, will proceed ahead to visualizing using Lag plots. Lag plots are used to determine whether there is a random time series or dataset. There should be no common structure in "random data" in "lag plot". In the "lag plot", the uncommon structure is the sign of non-random data. Then will proceed with time resampling step. This step consists of variations in incidence of observations in time-series data. Resampling is done in two forms – "Upsampling" and "Downsampling". The rise in incidence of samples like "from minutes to seconds" is known as "Upsampling". On the other side, decline in incidence of samples like from "days to months" is known as "Downsampling". Next will perform time series decomposition & statistical tests which consists of "KPSS test" and "ADF test". KPSS test is known as Kwiatkowski-Phillips-Schmidt-Shin tests are used for experimenting with "null hypothesis" than stationary time series across the "stationary trend" over the unit root alternative. ADF test is known as Augmented Dickey-Fuller test. It is a widely used "statistical test" for testing stationary time series.

Feature Extraction would be the next step that would be done using rolling windows method. Next, would be to proceed with modelling the dataset. Will be using XG Boost, ARIMA, Prophet and LSTM methods for modelling. Once modelling of all the 4 techniques are done, then will compare those models among themselves. Post comparison of models will provide the best suited model among the four and start with the evaluation of the prices of cryptocurrency. For evaluation metrics, will be using “Mean absolute error”, “Mean squared error” and “R^2 score”. Then will conclude the research by providing the best prices of the cryptocurrency with the help of the evaluation metrics used “(Cryptocurrency price forecasting based on short-term trend KNN model, 2021)”.

6.5 Model Building:

In this research, will be to perform different model building techniques like XG Boost, ARIMA, Facebook Prophet and LSTM. The models that will be built and later be compared among each other are as below: -

6.5.1 ARIMA

The abbreviation of ARIMA is “Autoregressive Integrated Moving Average”. This machine learning model covers a range of temporal structures given in time series information. It consists of important elements of machine learning model itself, such as –

- “Auto Regression (AR)” which relies on dependent relation between some “lagged observations” and an observation.
- “Integrated (I)” which consists of differentiation of raw data like subtraction of any observation from the same of last time step to come up with stationary time series.
- “Moving Average (MA)” relies on dependency between a “residual error” and “observation” from the applicable “moving average model” for “lagged observations”. A standard “ARIMA (p, d, q)” notation is used where integer values substitute the parameters to show the use of the given ARIMA model.

p: It refers to “lag observations” count in the model. It is simply known as “lag order”.

q: Also known as “order of moving average”, it refers to the size of moving average.

d: It refers to the differentiation of raw observations at various times. It also known as the “degree of differencing”.

ARIMA results in “stochastic time series” on weekly basis in the form of two different polynomials. The “auto-regression” is the first and “moving average” is the second step for such polynomials.

6.5.2 Prophet

Prophet is a machine learning model for “time series forecasting” on the basis of additive model where “non-linear trends” are fit with daily, annual, and weekly holiday effects and seasons. It is ideal with strong seasonal effects in time series and historical information of various seasons. It is ideal for changing trends, missing data, and outliers. It has four important elements as additive regression program. It looks for changing trends with the selection of “change-points”.

In case a forecasting model is unable to work properly, it is important to balance the method as per the existing issue. It takes proper knowledge of the overall workings of “underlying time series models” to tune those methods. For example, the “maximum orders” of differencing, “moving average components” and “auto-regressive components” are the first parameters of input to “automated ARIMA”. It is not in capability of a common analyst to make adjustment of orders to get away from the behavior and this kind of expertise is not easy to scale and gain. This way, the “Prophet package” adds to easily tunable “intuitive parameters”. It can easily be used by someone who don’t have experience in “forecasting models” to make right predictions for different issues in corporate scenario.

Prophet is an “Open-source program” to predict the availability of “time series data” in “R” and “Python”. It is published by the core experts in data science from Facebook. It records changes in trend, massive outliers, and big data. In addition, it is not easy to estimate mixed information manually. There is also lack in flexibility in completely automatic techniques as they are inflexible to combine important assumptions. Furthermore, unique abilities of data science are needed for pure estimations and they are not simple. Hence, Prophet is designed to make top-quality predictions effortlessly. It is optimized to make corporate predictions on Facebook “(Prophetic Analysis of Bitcoin price using Machine Learning Approaches, 2021)”.

6.5.3 XG Boost

It is an “open-source library” of software delivering a normalizing “gradient boosting” outline for Java, C++, R, Python, Perl, and other programming languages. In 2011, it was initially proposed by “Carlos Guestrin” and “Tianqi Chen” and it constantly improved and upgraded in several follow-up studies by data scientists. It is an “extreme gradient boost” model based on “Boosting Tree” models. It enhances “second order by Taylor” with loss function and can use “parallel computer multithreading” autonomously in the CPU. It also has several techniques to avoid overfitting “(Prophetic Analysis of Bitcoin price using Machine Learning Approaches, *2021)”*. It is compatible to Windows, Mac, and Linux. It is programmed to deliver a “Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library”. It works on distributed frameworks for processing like “Apache Spark”, “Apache Hadoop”, “Dask”, and “Apache Flink” and single machine. It is the best machine learning model for “supervised learning” problems. There are various features used with training data to predict “target variable”. It is widely used to minimize “regularized objective” which includes penalty term to define complexity of the model and a “convex loss” function (based on difference in target and predicted outputs).

6.5.4 LSTM

"LSTM", or "Long Short-Term Memory", are a special type of "RNN" model that can read "long-term dependencies". They are exclusively programmed to avoid the challenge of “long-term dependency”. By default, they can remember data over the long run and they won’t have any issue in learning RNNs. All RNNs have a chain of constant neural network modules. The “Single tanh layer” is a very simple structure in this repeating module in typical RNNs.
They also don’t have issues like “exploding/vanishing of gradient descent”. Future values can easily be predicted with “time series forecasting” using LSTM based on sequential, earlier data. It makes “demand forecasters” more accurate for businesses to make decisions in a better way. It can easily triage the patterns of impact from various events “(Prediction of Bitcoin Price Using Bi-LSTM Network, 2021)”.

6.6 Evaluation Metrics

In this study, will later compare the models with each other and provide the suited technique to predict or forecast the prices of cryptocurrency. The sklearn metrics module implements several losses, score and utility functions to measure regression performance. We employ Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) methods to assess the conjecturing accuracy of these models (Prophetic Analysis of Bitcoin price using Machine Learning Approaches, 2021).

Some of these will be used as:

- **Root Mean Squared Error: -** It determines the risk metric "mean square error" which is consistent to the proposed value of "squared error". It is used to check how close estimates or forecasts are to actual values. Lower the RMSE, the closer is the forecast to actual. This is used as a model evaluation measure for regression models and the lower values indicates a better fit.
- **Mean Absolute Error: -** It determines another risk metric "mean absolute error" which is consistent to the proposed value of "L1-norm loss" or "absolute error loss". It determines the intensity of errors on average in some forecasts, without knowing their path. Consistent variables and their accuracy are measured with it.
- **R^2 score**, the coefficient of determination: - R-squared says how good your model fits the data. R-squared closer to 1.0 says that the model fits the data quite well, whereas closer 0 means that model isn't that good. R-squared can also be negative when the model just makes absurd predictions. More precise predictions have a smaller amount of error. R^2 is relevant in this context because it is a measure of the error. Lower R^2 values correspond to models with more error, which in turn produces predictions that are less precise.

7. Requirement of Resources

There are eight attributes in the dataset. Hence, there are some software programs needed in a PC or laptop dedicated for developers. The same system should also have conducted some academic projects with same volume of data successfully. For more processing power, “Google Colab Developer” environment is highly recommended.

7.1 Hardware Requirements

Here are some of the configurations required for a laptop or PC used by the developers –

- Intel i3 processor or higher @1.80 GHz,
- 4GB RAM or more
- Windows 8.1 (64bit) or higher variant

7.2 Software Requirements

The “development IDE” would be “Jupyter Notebook” in “Anaconda distribution library” and dataset would be coded in Python. For processing of data, “NumPy”, “scikit learn” and “Pandas” are some of the open-source libraries of Python which would be used. In addition, graphs would be plotted with “Seaborn” and “Matplotlib” libraries.

8. Research Plan

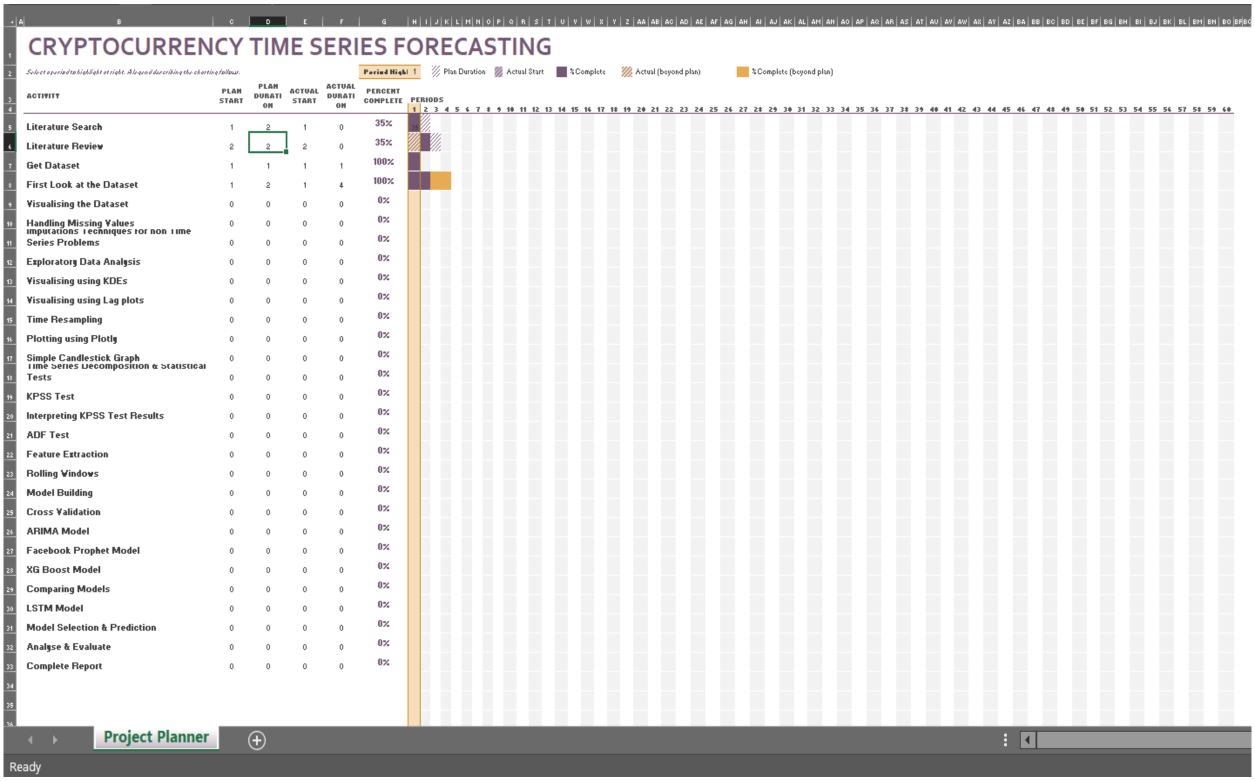


Figure 2: Project Plan Gantt Chart

References

C. Thomas *et al.*, Cryptocurrency Analysis using Machine Learning and Deep Learning, (*2021) IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2021, pp. 01-03, doi: 10.1109/SPMB52430.2021.9672268.

H. Jiang, Cryptocurrency price forecasting based on shortterm trend KNN model, (*2021) IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, 2021, pp. 1165-1169, doi: 10.1109/ICCASIT53235.2021.9633738.

V. Derbentsev, N. Datsenko, V. Babenko, O. Pushko and O. Pursky, Forecasting Cryptocurrency Prices Using Ensembles-Based Machine Learning Approach, (*2020) IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*, 2020, pp. 707-712, doi: 10.1109/PICST51311.2020.9468090.

N. Uras and M. Ortu, Investigation of Blockchain Cryptocurrencies' Price Movements Through Deep Learning: A Comparative Analysis, (*2021) IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2021, pp. 715-722, doi: 10.1109/SANER50967.2021.00091.

T. Phaladisailoed and T. Numnonda, Machine Learning Models Comparison for Bitcoin Price Prediction, (*2018) 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2018, pp. 506-511, doi: 10.1109/ICITEED.2018.8534911.

H. Kavitha, U. K. Sinha and S. S. Jain, Performance Evaluation of Machine Learning Algorithms for Bitcoin Price Prediction, (*2020) Fourth International Conference on Inventive Systems and Control (ICISC)*, 2020, pp. 110-114, doi: 10.1109/ICISC47916.2020.9171147.

A. Kovacs, B. Bogdandy and Z. Toth, Predict Stock Market Prices with Recurrent Neural Networks using NASDAQ Data Stream, (*2021) IEEE 15th International Symposium on Applied*

Computational Intelligence and Informatics (SACI), 2021, pp. 000449-000454, doi: 10.1109/SACI51354.2021.9465634.

N. P, R. J. Tom, P. Gupta, A. Shanthini, V. M. John and V. Sharma, Prediction of Bitcoin Price Using Bi-LSTM Network, (*2021) International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402427.

S. Pillai, D. Biyani, R. Motghare and D. Karia, Price Prediction and Notification System for cryptocurrency Share Market Trading, (*2021) International Conference on Communication information and Computing Technology (ICCICT)*, 2021, pp. 1-7, doi: 10.1109/ICCICT50803.2021.9510122.

S. Yogeshwaran, M. J. Kaur and P. Maheshwari, Project Based Learning: Predicting Bitcoin Prices using Deep Learning, (*2019) IEEE Global Engineering Education Conference (EDUCON)*, 2019, pp. 1449-1454, doi: 10.1109/EDUCON.2019.8725091.

T. Zhu, Z. Zhou, C. Liao, X. Li and Q. Zhang, Unsupervised Learning Based Detection Method for The Life Cycle of Virtual Cryptocurrency, (*2021) IEEE 6th International Conference on Smart Cloud (SmartCloud)*, 2021, pp. 44-50, doi: 10.1109/SmartCloud52277.2021.00015.
J. Sun, Y. Zhou and J. Lin, Using machine learning for cryptocurrency trading, (*2019) IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*, 2019, pp. 647-652, doi: 10.1109/ICPHYS.2019.8780358.

R. G. Tiwari, A. K. Agarwal, R. K. Kaushal and N. Kumar, Prophetic Analysis of Bitcoin price using Machine Learning Approaches, (*2021) 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, 2021, pp. 428-432, doi: 10.1109/ISPCC53510.2021.9609419.

Zielak, Bitcoin Historical Data: bitstampUSD_1-min_data_2012-01-01_to_2020-09-14.csv, [online] Available: https://www.kaggle.com/caesarlupum/bitcoin-price-prediction-auto-ts/data?select=bitstampUSD_1-min_data_2012-01-01_to_2020-09-14.csv