MEF UNIVERSITY

Predicting the Price of Bitcoin: Using Machine Learning Time Series Methods

Capstone Project

Sezer Ulutaş

Asst. Prof. Dr. Utku Koç

İSTANBUL,2020

MEF UNIVERSITY

Predicting the Price of Bitcoin: Using Machine Learning Time Series Methods

Capstone Project

Sezer Ulutaş

Asst. Prof. Dr. Utku Koç

İSTANBUL, 2020

MEF UNIVERSITY

Name of the project: Predicting the Price of Bitcoin
Name/Last Name of the Student: Sezer Ulutaş
Date of Thesis Defense: 14/12/2020

I hereby state that the graduation project prepared by Sezer Ulutaş has been completed under my supervision. I accept this work as a “Graduation Project”.

14/12/2020
Asst. Prof. Dr. Utku Koç

I hereby state that I have examined this graduation project by Sezer Ulutaş which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

14/12/2020

Director
of
Big Data Analytics Program
Prof. Dr. Ozgur Ozluk

We hereby state that we have held the graduation examination of Sezer Ulutaş and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Dr. Utku Koç	………………………..
2. Prof. Dr. Özgür Özlük	………………………..
3. ……………………	………………………..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Sezer Ulutaş 14.12.2020 Signature

EXECUTIVE SUMMARY

PREDICTING THE PRICE OF BITCOIN:
USING MACHINE LEARNING TIME SERIES METHODS

Sezer Ulutaş

Advisor: Asst. Prof. Dr. Utku Koç

DECEMBER, 2020, 39 Pages

Cryptocurrencies have greatly increased their Bitcoin-led popularity in recent years due to increased trading volumes and massive capitalization in the market. These cryptographic forms of money are not just utilized for exchanging nowadays, they are additionally acknowledged for fiscal exchanges. It appears to be evident that financial specialists, dealers and people, in general, are progressively intrigued by bitcoin and altcoins as costs rise and the arrival on ventures made increments. This examination centres around applying estimate models that will make precise value forecasts for cryptographic forms of money. The data were taken from two different exchanges and evaluated as combined dataset. As a result of the evaluation, it was determined that the prices were close to each other in terms of value and the data were combined. We obtained the daily time series data by determining the Bitcoin weighted price as a dependent variable and Open, Close, High, Low and Volume as independent variable. We predicted the next 6 months with ARIMA, LSTM and XGBoost methods. We compared these estimates using MSE, MAE, MAPE and R squared performance metrics. LSTM is the model with the best R squared value of 29.7%. In the process performed by taking the average of LSTM, XGBoost and ARIMA performed with the name of Average ML method, the R square value was found to be 41.6% as a much better result than LSTM.

Key Words: Time Series Analysis, Price Prediction, Cryptocurrency, XGBoost, LSTM, ARIMA, Various Boosting, Bitcoin

ÖZET

MAKİNE ÖĞRENİMİ ZAMAN SERİİSİ YÖNTEMLERİNİ KULLANARAK
BİTCOİN FIYATINI TAHMİN ETME

Sezer Ulutaş

Danışman: Dr. Öğr. Üyesi Utku Koç

ARALIK, 2020, 39 Sayfa

Kripto para birimleri, artan ticaret hacimleri ve piyasadaki büyük kapitalizasyon nedeniyle son yıllarda Bitcoin liderliğindeki popülerliğini büyük ölçüde arttırdı. Bu para birimleri bugünlerde sadece ticaret için kullanılmıyor, aynı zamanda parasal işlemler için de kabul ediliyor. Yatırımcılar, tüccarlar ve halkın fiyatlar yükseldikçe ve yapılan yatırımların geri dönüşü arttıkça Bitcoin ve altcoinlere giderek daha fazla ilgi duyduğu açık şekilde görünüyor. Bu araştırma, kripto para birimleri için doğru fiyat tahminleri yapacak tahmin modellerini uygulamaya odaklanıyor. Veriler iki farklı borsadan alınarak değerlendirilmiştir. Değerlendirme sonucunda fiyatların değer olarak birbirlerine yakın oldukları tespit edilmiş ve datalar birleştirilmiştir. Birleştirilen datalar projede "birleşik veri" ismi ile geçer. Bitcoin ağırlıklı fiyatını bağımlı değişken ve açılış, kapanış, en yüksek ve en düşük fiyatı bağımsız değişken olarak belirleyerek günlük zaman serisi verilerini aldık. Elde ettiğimiz sonuçları ARIMA, LSTM ve XGBoost yöntemleri ile bir sonraki 6 ayı tahminledik. Bu tahminleri MSE, MAE, MAPE ve R kare performans metriklerini kullanarak karşılaştırdık. LSTM %29,7 oranı ile aralarındaki en iyi R kare değerine sahip modeldir. Ortalama makine öğrenmesi yöntemi ismi ile LSTM, XGBoost ve ARIMA'nın ortalamaları alınarak gerçekleştirilmiş olan işlemde LSTM'den çok daha iyi bir sonuç olarak R kare değeri %41,6 olarak saptanmıştır.

Anahtar Kelimeler: Zaman Serisi Analizi, Fiyat Tahmini, Kripto Para Birimleri, XGBoost, LSTM, ARIMA, Yükseltme Yöntemleri, Bitcoin

TABLE OF CONTENTS

1. INTRODUCTION

Explorations have been done on exchanging volume and its relationship with valid statement returns. That connection between business stock prices and future prices is contrary with a powerless market proficiency (Campbell, 1996), exploring the relationship has pulled in the consideration of analysts and financial specialists. One reason why incredible significance is given to these connections is that developing values can bring adequate pay on the off chance that it is all in all correct to settle on the exchange volume.

Building strategies and frameworks in enormous information science is routinely utilized in money related market applications, including signal handling. Computerization of business sectors (Nakamoto, 2008) advances examination and algorithmic arrangements that are surely known and taken care of by designing networks. Moreover, the recent success of machine learning is noticeable coming out every day. Computer vision gained importance to improve the display of money-related markets. Machine learning led to consideration of the monetary network that was looking for effective procedures like regular language. In this project, we are looking for a solution to the problem of benefit allocation. On the other hand, we are trying to cascade dynamic problems through direct cooperation with the world. As a result, in this project, we explore how it can be solved with the most versatile machine learning.

Next, we provide more information about blockchain and crypto currencies including the history of bitcoin - the first crypto currency. We then provide a literature survey of the techniques used in this study.

It should be noted that even if everything is considered and subjected to a model, each technique can behave differently under various conditions.

1.1. Blockchain

Blockchain is a distributed data recording system that allows tracking of sequential transactions. It should not be confused with a database due to its structure. Transactions and recorded data on the blockchain cannot be changed or deleted. It owes this feature to the fact that it was first distributed across different devices. The next important feature is that it connects the blocks where the data is collected with encryption algorithms just like a chain. Thanks to the public keys and private keys inside, it also prevents anyone else from accessing important files. Blockchain enables us to keep records that cannot be changed and manipulated with this key structure. When it is attempted to be replaced, it dismisses blockchain devices that perform erroneous transactions or manipulations from the network. The main thing that makes this technology is that it doesn't need a central authority. Transactions are recorded in the ledger and fully democratically propagated by computers on the network. The more computers that join this network, the greater the reliability of this system. Blockchain's capabilities are not limited to the financial sector. Studies on this issue are continuing in many other sectors.

All entities such as the manufacturing sector, notary operations, financial services, GDPR (General Data Protection Regulation), accurate registration of written publications, IOT (Internet of Things) can have blockchain technologies in themselves. An ecosystem that is desired to be used with more than one device can be easily created with blockchain. Blockchain technology stores and protects the information of asset owners and those authorized to use or view them by protecting assets. Many transactions made in notaries with smart contracts are made on the blockchain and can be stored forever. For example, consider a cottage connected to the blockchain and the key lock of this cottage is connected to an IOT device. Someone who will settle abroad also wants to open an advertisement and sell his property for a certain amount. The important point here is that thanks to the smart contract opened, the payer is now the main owner of the new key after making the payment. This eliminates all intermediaries in between, providing a much faster and more reliable way.

1.2. Cryptocurrencies and Bitcoin

A crypto money (or crypto asset) is a digital asset, a virtual element that uses cryptography to secure its transactions, the way it works is designed as an alternative to cash exchange tool. Crypto currencies are a kind of digital currency, alternative currency and virtual currency. Unlike central electronic money and centralized banking systems, crypto assets are completely decentralized. The decentralization of each cryptocurrency comes from a blockchain that functions as a distributed ledger, the public transaction database. Satoshi Nakamoto published an article in 2008 called Bitcoin (the first decentralized cryptocurrency): A Peer-to-Peer Electronic Cash System (Nakamoto, 2008). Since then, many different crypto currencies have emerged. These are often called altcoins, which is a combination of alternative and coin. The emergence of bitcoin article was a crisis at that time and traditional exchanges with balloons formed in the event of unlimited assets. It was created on the bitcoin blockchain network, a limited, uncontrollable asset. We can briefly define Blockchain as a distributed database that provides encrypted transaction tracking. As the name suggests, Blockchain technology, built with a chained model, is traceable but difficult to break, allowing transactions without being connected to a center. Despite such developments, bitcoin still had no value. Adding thousands of added value to its value over the years, bitcoin has also been defined as the main value layer of a huge network by generating different cryptocurrencies. The stock market is the result of the transactions performed in certain periods turn into financial behavior on aboard. Everybody can view transactions in seconds, minutes, hourly, daily, monthly or even yearly. We try to make predictions on the prices that rise and fall in certain periods. What we want to do in this project is to see how these periods will move and to make transactions on certain values.

1.3. Chronologic of Bitcoin Prices

In August 2008, the domain name bitcoin.org was registered. (Bitcoinwiki, 2020) In May 2010, a person who was living in Florida opened 10 thousand bitcoin title for 2-person pizza on bitcointalk.org. Afterwards, someone sent a 2-person pizza from London to Florida and the first valuation was made. The swap scale on MtGox exchange was $ 0.50 per BTC, while the bitcoin share capital reached $ 1 million in November 2010. In June 2011, the USD / BTC rate was 10 USD. For 6 days, Bitcoin value stabilized at $ 31.91 on MtGox exchange, after which it remained volatile. In February 2013, the Bitcoin swap scale outperformed $ 31.91 over the previous 601 days. In April 2013, Bitcoin's exchange fees increased from 100 USD to 1 BTC. In January 2015, Coinbase raised a historic record for a Bitcoin organization, raising $ 75 million as a Series C investment. In February 2015, the Bitcoin price increased to $ 262. In January 2017, Bitcoin's price broke the US $ 1,000 record. In June 2017, the Bitcoin swap scale exceeded $ 3000. In November 2017, Bitcoin price exceeded $ 10,000. In December 2017, Bitcoin reached a record high, barely reaching $ 19,500. In December 2017, the price of bitcoins fell after South Korea announced extra measures to steer the bitcoin exchange. It dropped as low as $ 6,300 in October 2018. In August 2020, the price of bitcoin hovers between $ 11,000 and $ 12,000.

1.4. Literature Survey for LSTM

Long Short Term Memory is advanced recurrent neural network and it serves to see long term conditions in artificial neural networks, naturally to acquire "context" aware neural networks. Basically, by using a more complex flow mechanism between nodes, it eliminates the problem of "exponential growth of errors" caused by models such as backpropagation. For example, deep mind pictures are usually the work of this model. Adrian Et Al. (Ţăran, 2011) published an important article on this subject. The examination, investigating the consequences of the exploration, reasons that the outrageous qualities of RSI (Relative strength index) don't show the inversion of a pattern, yet that its heading proceeds, in any event for the time being. The relative strength index is a technical indicator used in the analysis of financial markets. It was developed by J. Welles Wilder in 1978 and was first published in his book. RSI is an indicator of momentum and simply generates overbought and oversold signals. In this way, the old way is pointless as the converse translation yields positive outcomes for the two types of the pointer. Suggested technique, Hari Et Al. (Yulius, 2017), is using the framework that enables showing the purchase in three categories: sell, oversold, and unbiased time. In this manner, it can enable the financial specialist to reflect and demonstrate the dynamic procedure. Also, the framework can help with breaking down verifiable value information for the chose timespan. The time period may likewise influence the result of the gauge, particularly the moving normal outcome. In view of client experience study, usability in the framework is worthy. Finally, the framework can't decide if it is a misfortune or benefit, it can only show an ideal opportunity to purchase or sell where the choice is up to the merchant (Talreja, 2017). The idea of RSI inversion applies to record-breaking outlines and gives an away from of market slants just as estimating market headings. This idea of relative quality is valuable not exclusively to feature promising speculations, yet in addition to empowering you to pass judgment on your own exhibition. While there are different relative reinforcing measures accessible for financial specialists, they all look to hold onto stocks that show more grounded value activity than those exhibited by the overall market. Like all venture procedures requires cautious examination and comprehension of the elements that trigger stock value activity. Bhargavi (2017) introduced an article in which they clarified that RSI is one of the best-specialized examination apparatuses accessible and can be utilized successfully to fabricate a portfolio. Also, the P/E (Price to Earnings) proportion has been found to all the more likely mirror the presentation of an association contrasted with EPS (Earnings per Share). In spite of the fact that RSI is an exceptionally amazing scientific instrument in itself, utilizing central investigation and other specialized logical apparatuses together gives better outcomes. In the article introduced by Hochreiter Et Al. (Hochreiter, 1997) gave data about LSTM and its design just as the numerical understandings that every cell in the LSTM conveys in its procedure. They even depicted the steady blunder stream inside the CEC (Commodities Exchange Center) and strategies for decreasing the mistake by utilizing two entryways at the information and yield stream of the mistake. Subsequently, they arrived at the resolution that their models will guarantee LSTM's progression of information in every memory cell without mistakes. Sang Et Al. (2019) distributed and proposed the long short term memory for stock expectation. It makes it harder for a normal broker to make a benefit as will need to spend many days to decide, whereas the stock forecast just relies upon development of past values and will require much less time. The writer adds that it needs to be observing the stock pattern and making an important move.

1.5. Literature Survey for XGBoost

There has been an incredible discussion in securities exchange expectation, particularly Machine Learning. Expectation of securities exchange conduct is utilized to make a benefit by recognizing future patterns. Forecast is normally made by examining verifiable time arrangement information. For example, Logistic Regression, SVM, Linear

Regression, KNN, NN and Linear Discriminant Analysis are utilized in stock estimation. (Chen, 2018) Strategic relapse was seen as truly outstanding with a triumph pace of 55.6%, and examination indicated that the accuracy rate was higher than different models. On Machine Learning in Stock Price Trend Forecasting paper writers utilized preparing information from 3M Stock information (Dai, 2013). The information incorporates every day stock data going from 1/9/2008 to 11/8/2013 (1471 information focuses). These calculations incorporate SVM, Logistic Regression, Quadratic Discrimination Analysis. The calculations are applied to the following day model that predicts the consequence of the stock value the following day, and to the drawn out model that predicts the stock price for the following n days. The following day the forecast model delivered precision results running from 44.5% to 58.2%. SVM revealed the most noteworthy precision at 79.3% for the drawn out estimate, where the time period was 44. In one of the articles (Imandoust, 2014) distributed on ANN (Artifical Neural Network) utilized to anticipate the Japanese bearing. Trade. It gave 81.2% accuracy (Timmermann, 2019).

1.6. Literature Survey for ARIMA

Tansel et al. (Tansel, 1999) analyzed the performance of direct optimization, ANNs, and hereditary algorithms in displaying time arrangement information dependent on demonstrating precision, accommodation, and calculation time. The examination uncovered that direct streamlining methods give the best estimates with hereditary calculations that give comparative outcomes when the restrictions of boundaries and goal are carefully chosen, while Neural Networks give the most exceedingly awful estimates. The examination, revealed in (Lee, 2007), additionally showed the forecasting performance of ARIMA and ANN models in the figure at the Korean Stock Price Index. The ARIMA model gave more accurate expectations than the utilized backpropagation neural network. This is more articulated for the middle of the road figure skylines.

2. ABOUT THE DATA

Bitcoin datasets are collected with API from Coinbase (Coinbase, 2020) and Bitstamp (Bitstamp, 2020) Coinbase and Bitstamp datasets are the open sources of data collected and researched by relevant companies so researchers can conduct their research. Bitcoin dataset includes the minute duration, open price, the most valuable price within that minute, the least valuable price within that minute, the close price, volume, volume within that day, and weighted price information. Combined dataset represents the combined version of bitstamp and coinbase datasets.

We brought together two different bitcoin datasets and reassembled these datasets. Bitstamp data set consists of 4363457 lines and coinbase 2099760 lines. The combined dataset had more than 6 million rows.

Coinbase and Bitstamp data presents the conjuncture of the general condition of prices over the years in the cryptocurrency market. The purpose of using this data is to obtain information about whether the specified machine learning methods have a significant impact on the cryptocurrency market. Data were taken from 2 different exchanges and evaluated. As a result of the evaluation, it was determined that the prices were close to each other in terms of value and the data were combined. We did that because every cryptocurrency market has its own dynamics. Every exchange opens or closes a trade on its own prices on their own community. On one side, the price may rise by $10, in some cases it may be $10 lower on the other.

2.1. Exploratory Data Analysis

Timestamps with no exchanges or movement have their information fields loaded up with NaNs. In the event that a timestamp is missing or bounces, it might be on the grounds that the change (or API) is shut, the change (or API) isn't accessible, or another unexpected specialized mistake in information revealing or assortment. Each exertion has been made to deduplicate the sections and confirm that the substance is right and complete in the most ideal manner.

Approximately 10069 values had a NaN value. I used an operation method to avoid any gaps when plotting the NaN values in the Open, High, Low, Close and Weighted_Price values lines. Using the filling method, I was subjected to further filling and replaced the last non-null value observed until I encountered another non-null value. The reason is to show that the current data looks the same in time series without distorting its trend and to complete the training since machine learning will be used.

On the other hand, Volume_(BTC) and Volume_(Currency) values had NaN values. We changed Volume_(BTC) and Volume_(Currency) values with 0. Blank columns cleaned from the data frame.

Figure 1 shows historical bitcoin prices daily. We wanted to review the open, close and weighted avg values and check how different they are from each other.

Figure 1: Historical Bitcoin Prices Daily

Figure 2 shows historical bitcoin volume experienced every week. What I want to see here is whether the effect reflected on the price is seen as the volume increases. When the price table and this table are placed side by side, it is seen that the volume may be a correlation with the price for some dates.

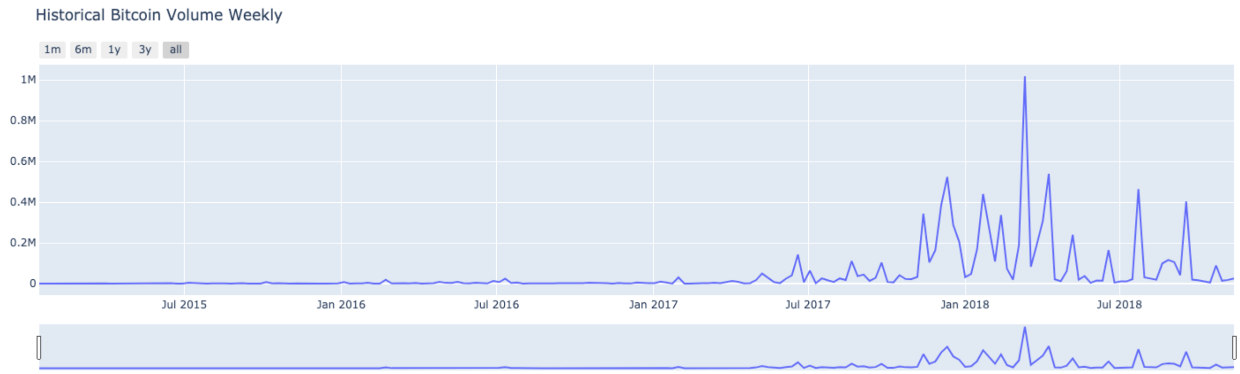

Figure 2: Historical Bitcoin Volume Weekly

Figure 3 shows the representation of bitcoin volume and weighted USD on a graph. As mentioned above, this may be true for some prices and volumes when the charts are combined.

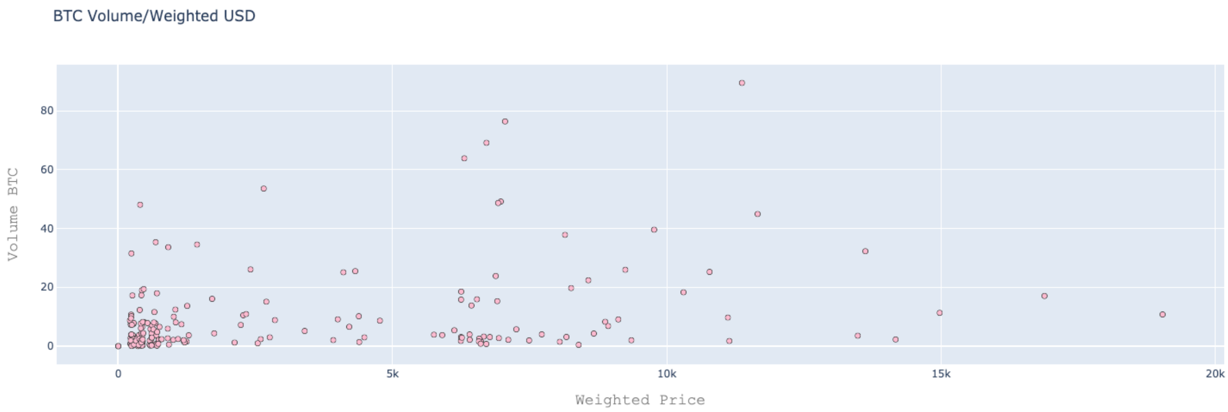

Figure 3: BTC Volume/Weighted USD

Figure 4 shows separation between train data and test data daily. The orange painted part denotes the training set. The blue colored part refers to the test set.

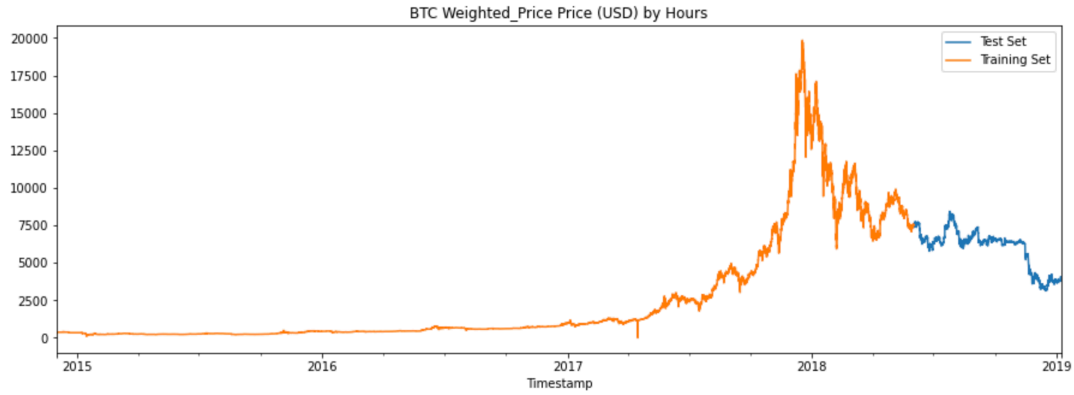

Figure 4: Daily Training Dataset and Test Dataset

We wanted to check correlation between the columns and below Figure 5 shows correlation of columns Volume is correlated to Weighted Price and Weighted Price is directly related to Open, High, Low and close.

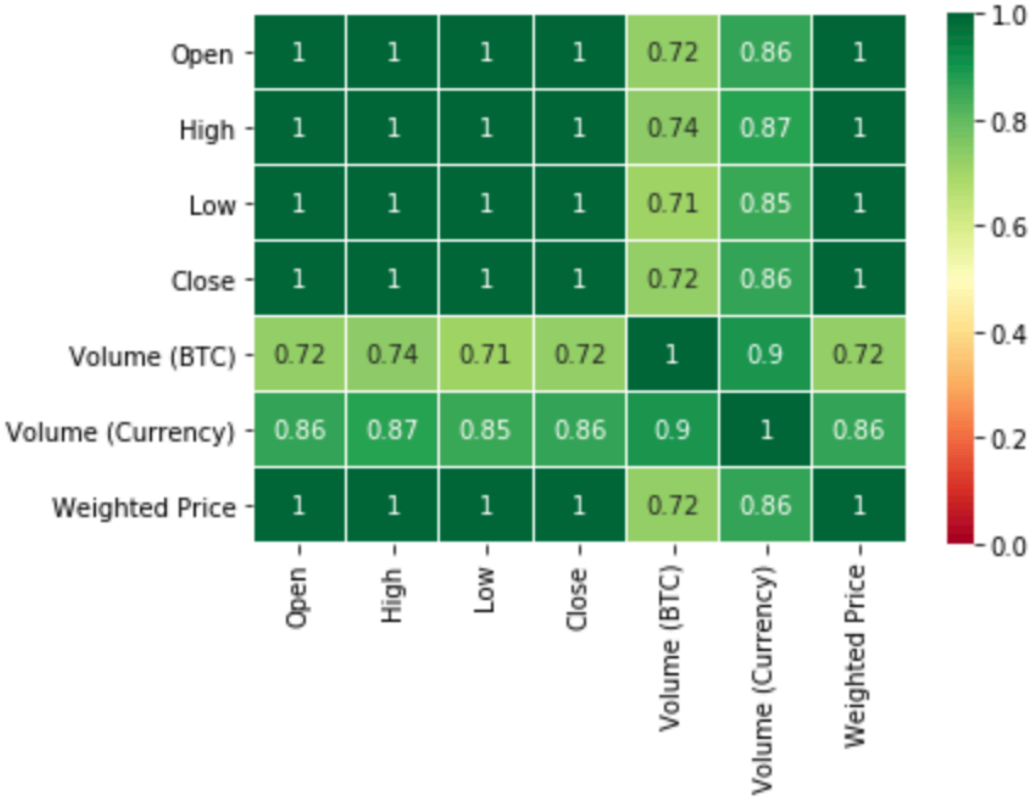

Figure 5: Heatmap of the dataset

3. PROJECT DEFINITION

3.1. Problem Statement

Changes are an intriguing part of human progress. The challenges in demonstrating the behavior of money-related markets, such as shaky, poorly predictive practices and poor chronicized coordination, have attracted mainstream researchers over the previous 50 years and attempted to use various design strategies. (Schinckus, 2017). Traditionally, the scientific details of the dynamic frameworks can be found in the book of John G. Proakis called Signal Processing and Control. The investigation offers positive help to demonstrate the widespread use of artificial intelligence techniques as (multiple crypto trading) exchange experts. Despite financial interchange and the information, they prepare, they can be filled as speculation procedures between sources and markets, which can reduce computational and memory complexity. The feedback on digital currency showcase information is addressed by the increase in knowledge (a useful methodology) and the progress of validated provisioning methodologies based on current metrics. Execution points of interest are accepted with the revitalized advertising arrangement (data-based) up to real market information. With more precise trading, an automated guide was needed to increase utility. Automatic guidance is not included in this project. In this capstone project, it is aimed to measure the performance of the models by trying to create a price estimation regarding the rising and falling bitcoin prices in certain periods.

3.2. Project Objectives

This project aims to estimate price by using machine learning models with a specific approach by looking at the price movements of bitcoin. Then, a model whose performance is improved is selected by applying and doing the necessary steps.

3.3. Project Scope

The price estimation shown in project analyzed with a long-term perspective. In other words, it does not include features such as processing prices in real time and turning them into a profit model. It tests whether prices can be predicted by looking at historical data. So, the fluctuation of prices may be the reason for many developments in the relevant dates and will be discussed separately in the future works section. A long-term date range has been used to emphasize a general pricing estimate.

3.4. Methods, Tools, and Techniques

The previous stages of selecting data are preliminary to discover the details of the data with Python 3.7. Afterwards, preprocessing steps are completed using Jupyter Notebook. Final recovery, modeling and hyperparameter adjustment processes are also performed on Jupyter Notebook.

The dataset used in this study was obtained by extracting bitstamp and coinbase data via API. By comparing the data with each other, it has been checked whether there is a different pricing than the general bitcoin pricing. The timestamp data contained in it is converted to date in the process of processing the data and modeling continues over date. From the data obtained as NaN, 'Open', 'High', 'Low', 'Close' and 'Weighted_Price' values are replaced with the last observed value and 'Volume_ (BTC)' and 'Volume_ (Currency) are replaced with the zero value. 'Since it is desired to start from a certain time interval, data begins to be processed from the beginning of 2015 and this processing continues until the 9th of January 2019. The reason for choosing this date is that the model is desired to be affected as little as possible from the price fluctuations in bitcoin. Starting at $300, as clearly seen in the data, pricing increases to $ 19,000 in December 2015. By making a correction from there, the price will advance from the $6,500 level by the end of 2018.

The most important item in determining the test data set and training the data set is the correct acquisition of the data to be used. According to the calculations made, this is reserved as training dataset for time series data from 2015 to 1 June 2018 and test dataset for the rest.

After preparing the data for modeling, the LSTM and XGBoost classifier models are equipped with best settings for prediction accordingly backtests.

3.4.1. ARIMA

Auto Regressive Integrated Moving Average model is used to look at recorded values and then predict what new future values will be based on the recorded values. In comparison, Auto Regressive and Moving Average processed for shaping ARIMA (Karakoyun, 2018) model. Regulation and determination is a method that should always be applied for ARIMA. For this model, it is desired to use finite differentiation to make the information more stable for a variable time series.

If we explain the parts of ARIMA, auto-regressive part means a pattern of growth/decline in the data is accounted for. Intagrated part means the rate of change of the growth/decline in the data is accounted for. Moving average means noise between consecutive tiem points is accounted for.

A value Y at time point t and adding/subtracting based on the Y values at previous time points (e.g., t-1, t-2, etc.), and also adding/subtracting error terms from previous time points. The formula of ARIMA is $Yt=c+\phi 1ydt-1+\phi pydt-p+...+\theta 1et-1+\theta qet-q+et$ where "e" is an error term and "c" is a constant.

ARIMA models are typically expressed "ARIMA(p,d,q)". Meaning of the "p" is the number of preceding (lagged) Y values that have to be added/subtracted to Y in the model, so as to make better predictions based on local periods of growth/decline in our data. Meaning of the "d" is the number of times that the data have to be "differenced" to produce a stationary signal. And lastly meaning of the "q" is the number of preceding/lagged values for the error term that are added/subtracted to Y.

3.4.1.1. Transformation for Time Series

A transformation can be useful if the data show increasing or decreasing variation over time. For example, the logarithmic transformation is a useful method of use in time series. Logarithms are useful because they are interpretable: changes in a log value are relative (or percentage) changes on the original scale. So, if log base 10 is used, then an increase of 1 on the log scale corresponds to a multiplication of 10 on the original scale.

Another useful feature of log transformation is that they constrain the forecasts to stay positive on the original scale.

3.4.1.2. Log Transformation

We used simple log parameter in python 3.7. Which means log(a).

3.4.1.3. Explanation and Mathematical Formula of Box-Cox Transformation

A box cox transformation is a transformation of a non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if our data isn't normal, applying a Box-Cox means that we are able to run a broader number of tests.

In mathematical term $w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$ is the formula of box cox transformation.

3.4.1.4. Lags of Time Series

When the time base is shifted by a given number of periods, a lag of time series is created. Lags of a time series are often used as explanatory variables to model the actual time series itself. The underlying reasoning is that the state of the time series few periods back may still has an influence on the series current state.

3.4.2. LSTM

Long Short Term Memory models are an extraordinary recurrent neural network equipped for getting data about long haul conditions. This implies they can recollect data longer than a RNN (Recurrent Neural Network). This profound learning model takes into consideration the preparation of enormous designs and is better for mistake opposition. (Saxena, 2018) This model, created by Hochreiter and Schmidhuber (Hochreiter, 1997) has arrangement of numerious layers including usable dataset and with at any rate one covered layers. It continues learning in various methods for time and involves recall and neglect entryways that help choose the information to be passed on as demonstrated by its criticalness and power. (Karakoyun, 2018)

3.4.2.1. Structure of LSTM Model Description

At the beginning of the graph as we can see on the figure 6 h_t , C_t means hidden layer vectors. After that step x_t comes directly to h_t side as Input vectors. It processed b_f, b_i, b_c, b_o named bias vectors and W_f, W_i, W_c, W_o comes after that means parameter matrices. At last stage σ and $\tanh$ comes to the process named activation functions.

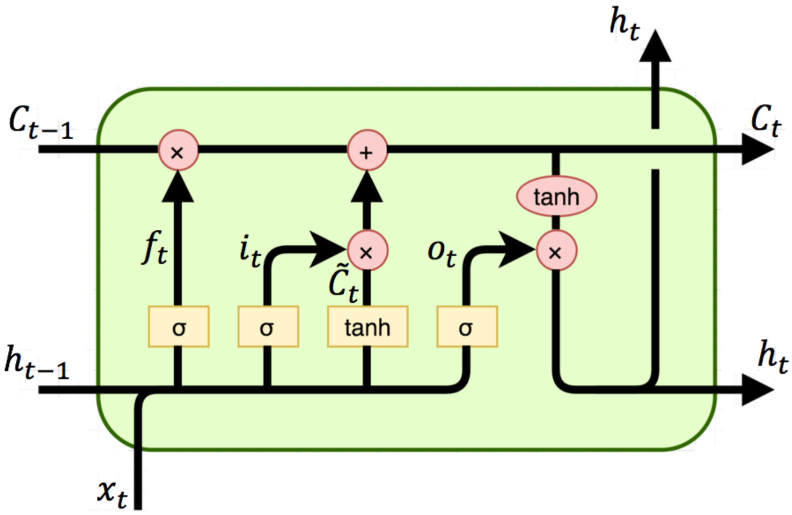

Figure 6: Structure of LSTM Model

As shown in Figure 6, the input gate, the forget gate, and the output gate of LSTM are designed to control what information should be forgotten, remembered, and updated. Gating is a method to selectively pass the information that is needed. It consists of a sigmoid function and an element-wise multiplication function. The output value is within [0, 1] to allow the multiplication to then happen to let information flow or not. It is considered good practice to initialize these gates to a value of 1, or close to 1, so as to not impair training at the beginning.

In mathematical term, $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$, $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$, $c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$, $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$ and $h_t = o_t \odot \tanh(c_t)$ are equations used in LSTM.

The information process of LSTM cell is described in mathematical terms. First, there is a need to forget old information, which involves the forget gate. The next step is to determine what new information needs to keep in memory with an input gate. From that, it is possible to update the old cell state, Ct−1, to the new cell state, Ct. Finally, it decides which information should be output to the layer above with an output gate.

3.4.3. XGBoost

XGBoost represents eXtreme Gradient Boosting. It was introduced by Tianqi Chen and developed with many engineers (Chen, 2018). It uses the gradient increment framework. It is a library that parallelizes the development of angle supported trees (Basak, 2019). It means to limit the time needed to develop trees and accelerate the way toward advancing, which settles on gradient boosting decision trees (GBDTs) commonsense to utilize. A GBDT is a classifier that joins the aftereffects of numerous frail classifiers to make a solid forecast. It is an improved form of a decision tree in light of the fact that each tree is approximated by an enormous number of relapse capacities fi(x). By attempting to all the more likely arrange the residuals in the past tree, the blunder in characterization can diminish progressively. When each tree has been ideally approximated, the structure's scores and addition are determined to decide the best split. At last, the forecast aftereffect of the whole model is the aggregate of every choice tree. Like the arbitrary timberland, a subset of the highlights is utilized to construct each tree.

The mathematical algorithm (referred to as the model) makes predictions si based on trained data pi. For example, in a linear model, the prediction is based on a combination of weighted input features such as sˆi = ∑j θj pij. The parameters need to be learnt from the data. Usually, θ is used to represent the parameters and, depending on the dataset, there can be numerous parameters. The predict value si helps to classify the problem at hand, whether it be regression, classification, ranking, or others. The main motive is to find the appropriate parameters from the dataset used for training purposes. An objective function is set up initially which describes the model performance. It must be mentioned that each model can differ depending on which parameter is used. Suppose that there is a dataset in which “length” and “height” are features of the dataset. Therefore, on the same dataset numerous models can be set up, depending on which parameters are used.

3.4.4. Evaluation Metrics for Prediction

The model's forecasting performance was evaluated by means of the following measures: the mean squared error (MSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE) and R squared.

The MSE, which is an excellent and general-purpose error measure for numerical predictions that amplifies and penalizes large errors, can be expressed as follows formula seen the end of the sentence. $\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}$

Here, at is the actual value, Ft is the forecasted value, and n is the number of forecasted values. The MAE, which is used to measure how close predictions are to the eventual outcomes, can be expressed as follows $MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|$. Finally, the MAPE, which measures the predictive accuracy of a given forecasting method in statistics, can be expressed as follows $MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$. If $\bar{y}$ is the mean of the observed data $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ formula should be used. Then the variability of the data set can be measured with two sums of squares formulas:

The total sum of squares:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

The sum of squares of residuals:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

General definition of R squared:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

4. METHODOLOGY

The project implementation pipeline showed in Figure 5. Data is taken from Bitstamp (Bitstamp, 2020) and Coinbase (Coinbase, 2020) APIs. Bitcoin dataset then consolidated into a solitary information outline and the information is customized to time arrangement.

Forecast calculations and models are applied. The yield assessed dependent on the assessment standards for every one of the three models, and afterward the outcomes are contrasted with decide the best forecast calculation.

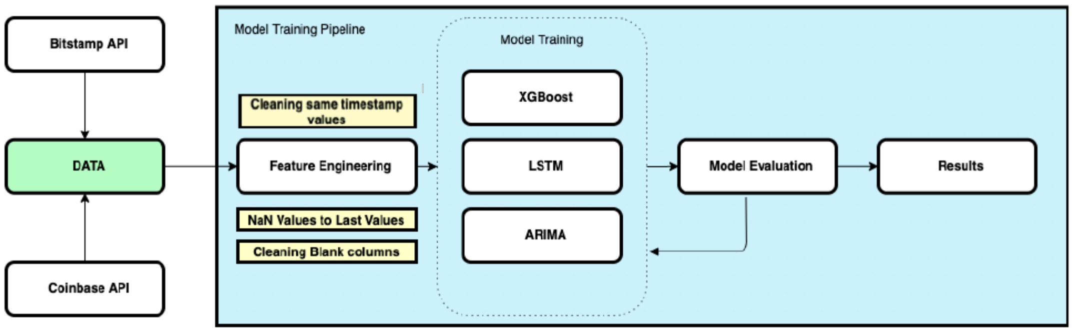

Figure 7: Project pipeline

Different datasets for Bitcoin can be gotten from various sources. In this manner, for this examination, information for Bitcoin was taken from the Coinbase and Bitstamp site utilizing the Bitstamp API and the Coinbase API. The library is utilized to serialize just as spare information, which gives us an information outline containing all separated information. Bitcoin trade information was recovered from Coinbase and imagined to check for any inconsistencies that may exist. NaN esteems, zero qualities and clear qualities were watched. It was noticed that finding the right Bitcoin prices are not simple, as the best approach to decide the prices are made by thinking about the interest and gracefully. Bitstamp and Coinbase are the two significant trades for Bitcoin.

4.1. ARIMA

The information was imagined, with a couple of irregularities watched, yet when I thought about the bitstamp and coinbase dataset, for the most part the prices were near one another. The information was cleared to dispose of the current zero qualities and the normal price was determined to get the genuine price to be utilized for investigation.

Invalid qualities were dissected after the information was cleared and gathered for bitcoin. Disclosure information examination was performed to all the more likely comprehend the investigation. The stats library has been integrated for ARIMA from the stats module on python 3.7. Value information were investigated to check for any patterns and irregularity that may be available. It has been seen that Bitcoin information contrast in change and mean over some undefined time frame. The necessary time frame cycles for this have been done with the help of pandas v1.1.2 library.

4.1.a. Original

We need to use time series analysis to make sense out of the given time series. We should note that our target variable was Weighted_Price. Firstly, defining a parsing function that uses seasonal decomposition to analyze the components of time series. The main purpose here is to measure whether the series is stationary or not. There are 2 things to do for this, first to perform seasonal trend decomposition and then to do a dicky fuller test. Our aim in seasonal trend decomposition is to see the components of time series and to obtain a result away from trends. Our aim in the Dicky fuller test is the null hypothesis that the analyzed time series are not stationary. In other words, the p value should be less than 0.05 during the hypothesis test. In such a case, our p-value is too high to reject the null hypothesis. The following plots were obtained as the output of the process.

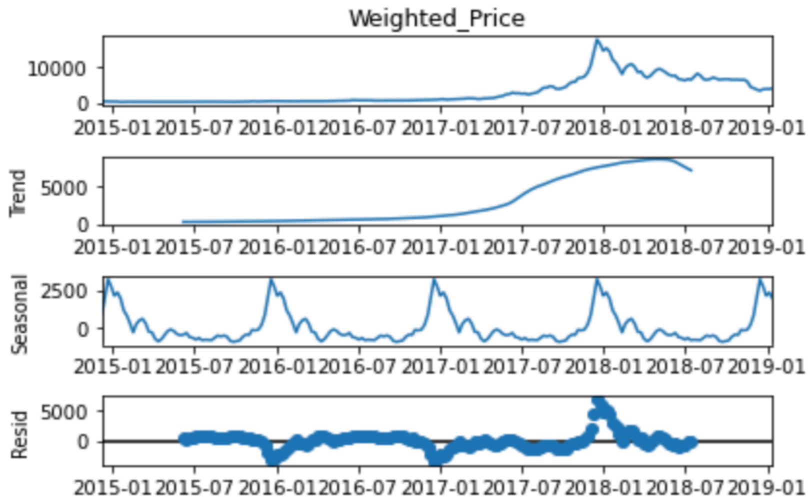

Figure 8: Trend, Seasonal and Resid Plots

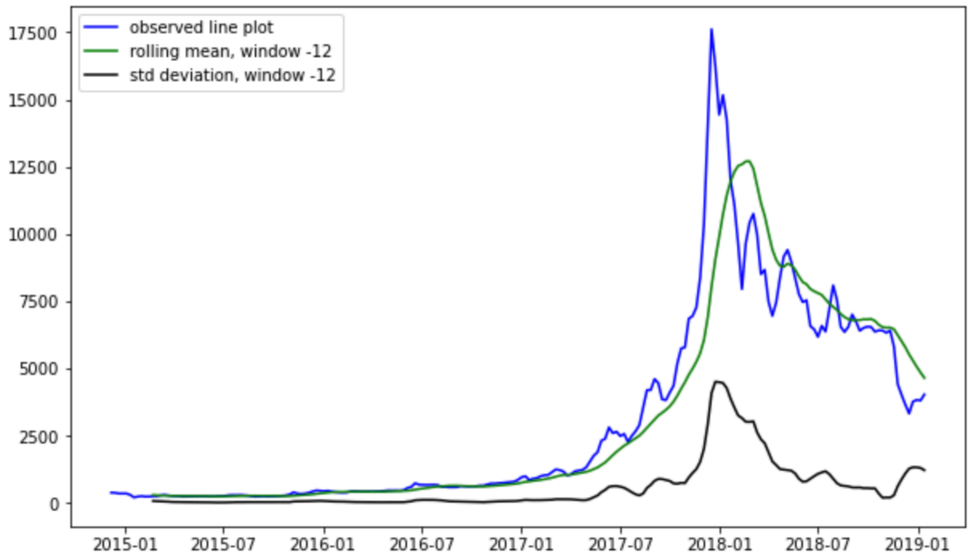

Figure 9: Observed Line Plot, Rolling mean and Std Deviation

4.1.b. Log Transformation

DF test p-value was found to be 0.529. With the visualization of the test and the p value, we can say that the series is not stationary. So it's early for us to apply the ARIMA model. We now consider a few transformations for our series and continue to check which one makes our series suitable for ARIMA modeling. By going through a series of transformations our series will first get a log of the price and the following plots were obtaine.

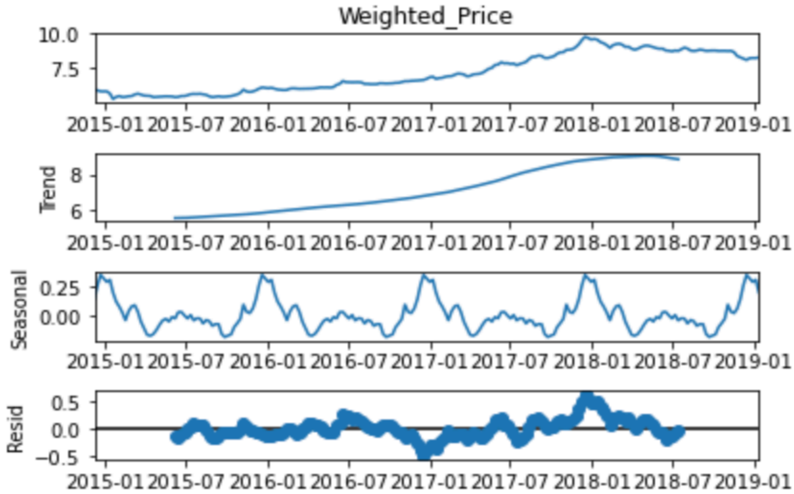

Figure 10: Log Transformed Trend, Seasonal and Resid Plots

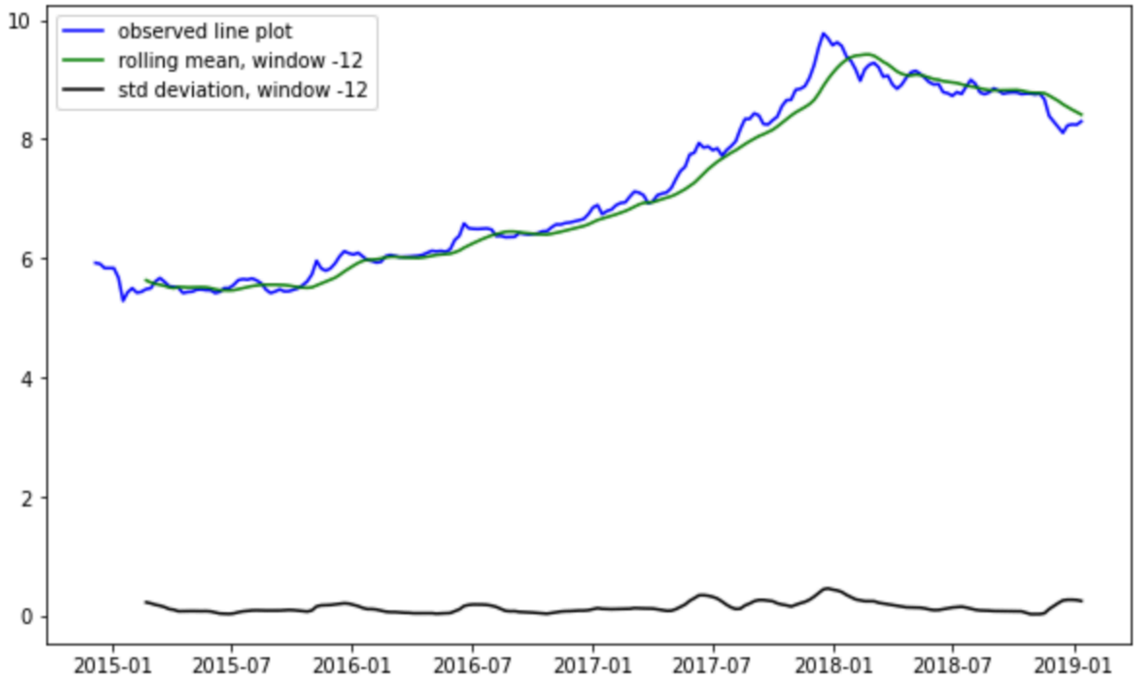

Figure 11: Log Transformed Observed Line Plot, Rolling mean and Std Deviation

4.1.c. Time Lag Applied Log Transformation

As we see in Figure 11 there is a lag between std deviation and Rolling mean and observed line plot. p-value was found to be 0.84. Based on the initial curve of the graph, the data needs to be transformed to get it stationary. First try transforming the data with log to remove the trend. Now we continue to apply the usual time lag operations applied to log converted prices and the following plots were obtained.

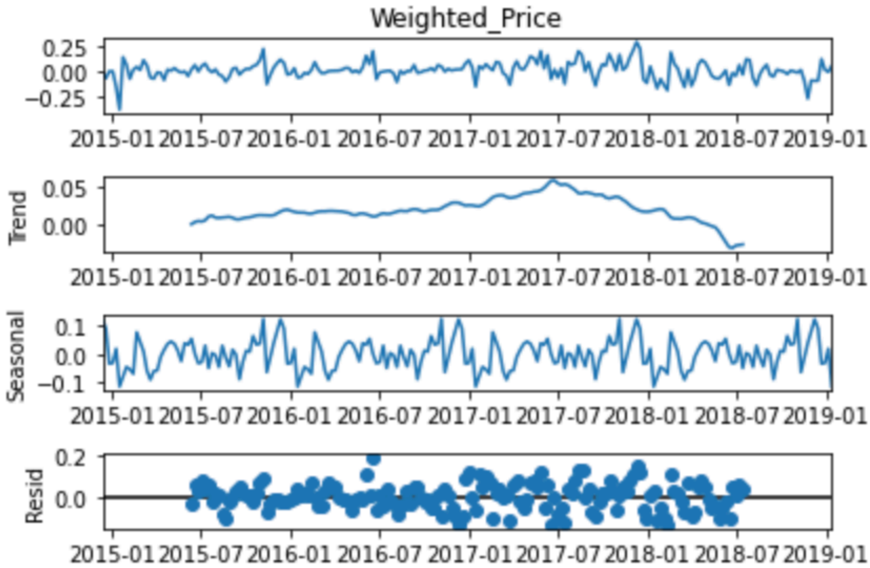

Figure 12 : Regular Time Lag Applied Log Transformed Trend, Seasonal and Resid Plots

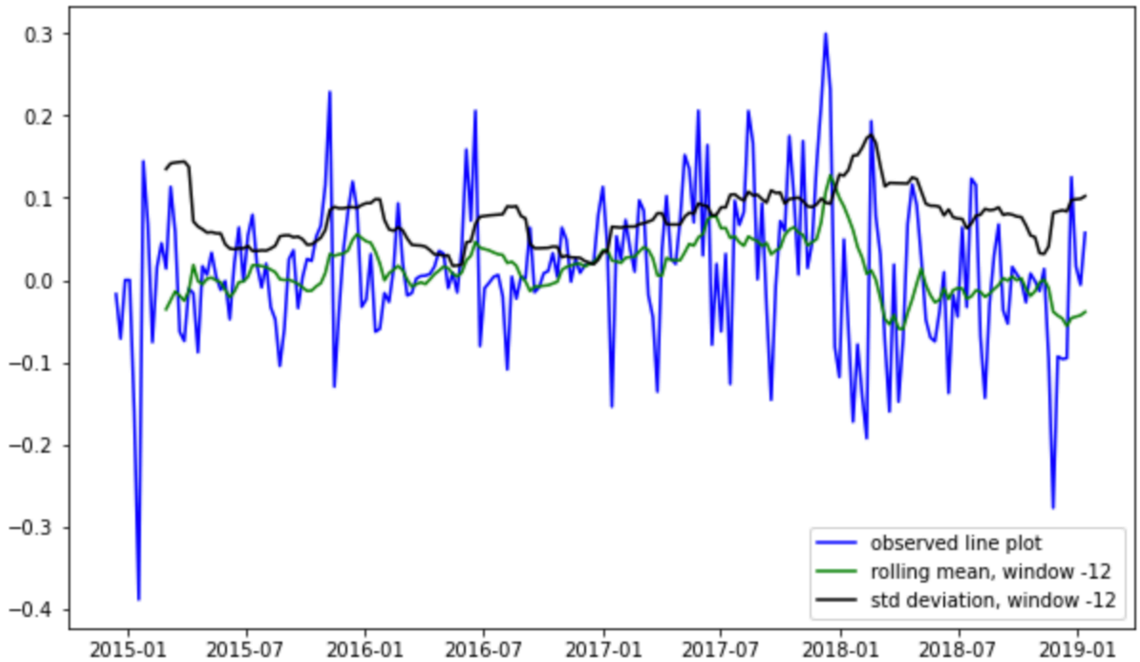

Figure 13: Regular Time Lag Applied Log Transformed Observed Line Plot, Rolling mean and Std Deviation

4.1.d. Box Cox Transformation

DF test p-value was found to be 0.00. Box cox changes then performed to smother fluctuation, trailed via seasonal separation and regular separation.

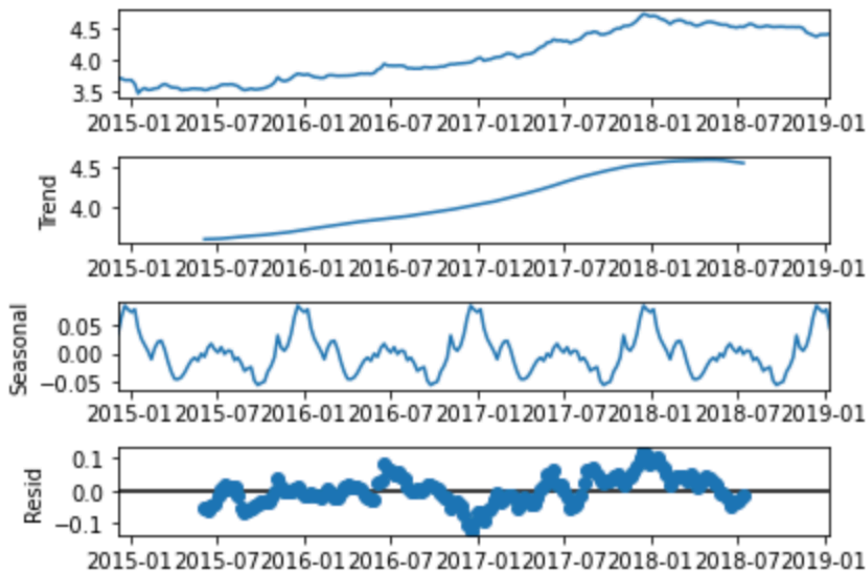

Figure 14: Box Cox Transformed Trend, Seasonal and Resid Plots

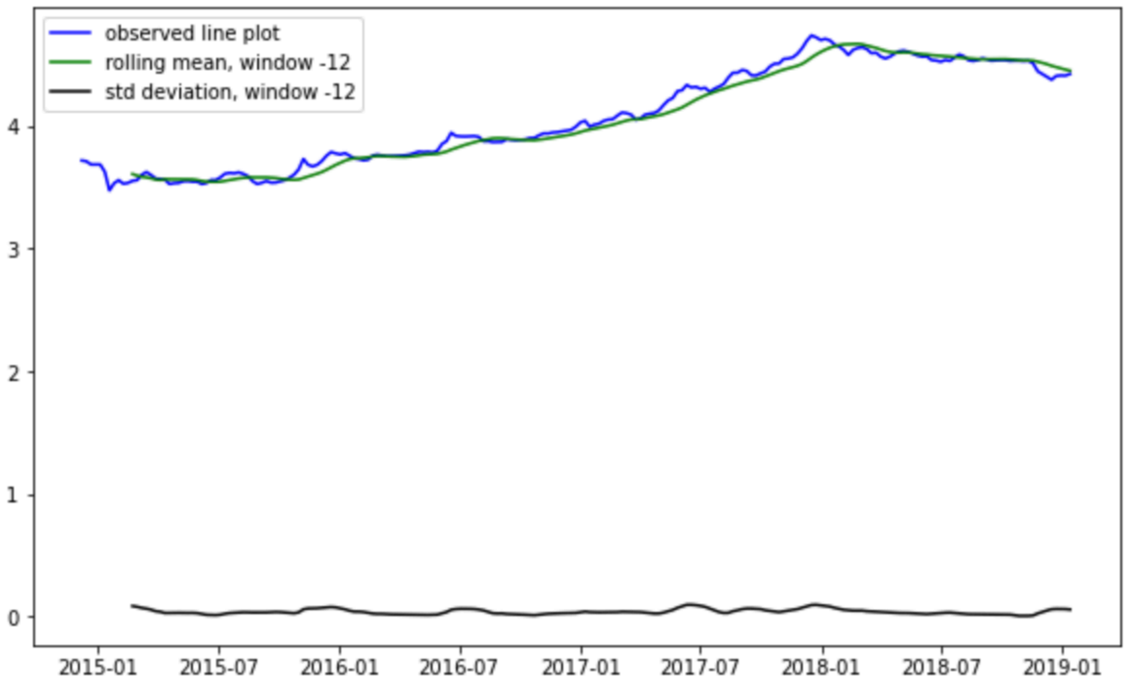

Figure 15: Box Cox Transformed Observed Line Plot, Rolling mean and Std Deviation

DF was obtained as test p-value: 0.76, lambda value: 0.12. Now we will examine the normal time lag applied to box cox transformed prices like the transactions applied above.

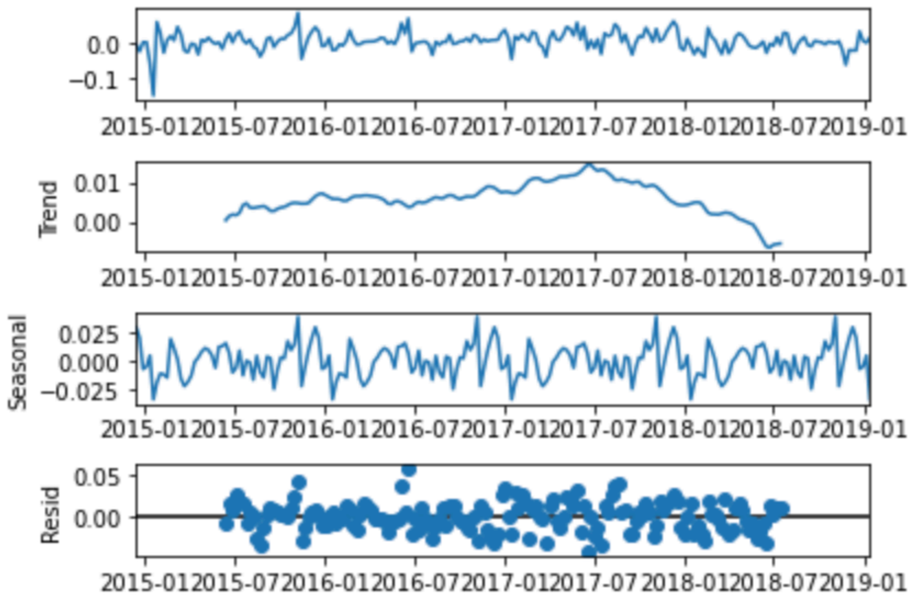

Figure 16: Box Cox Transformed Prices Applied Trend, Seasonal and Resid Plots

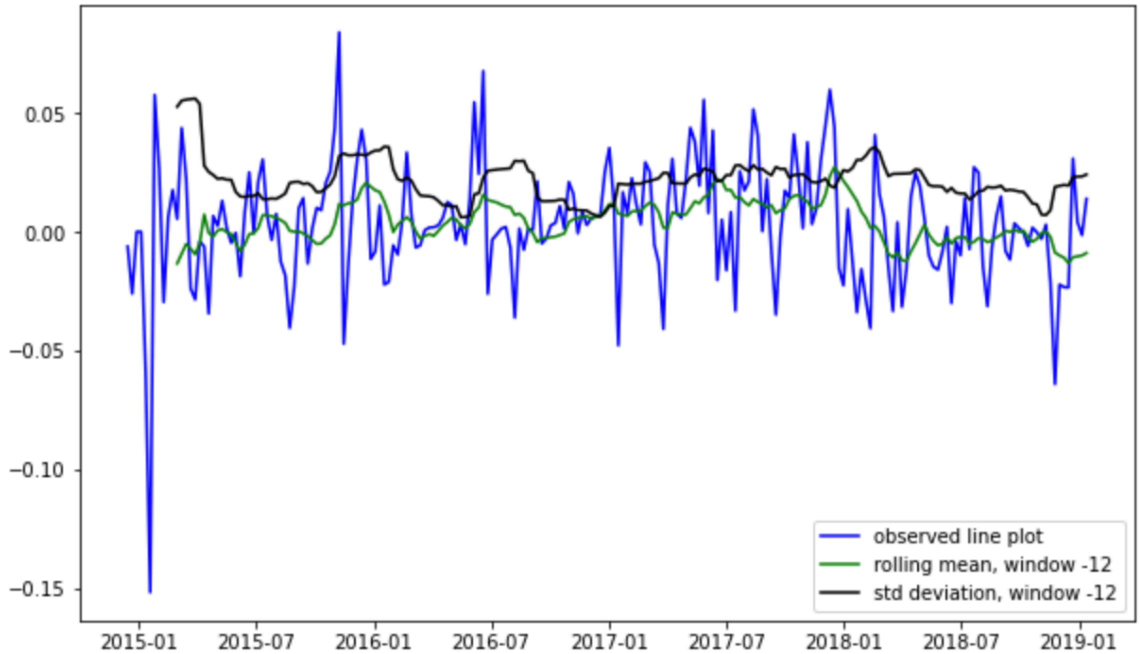

Figure 17: Box Cox Transformed Prices Applied Observed Line Plot, Rolling mean and Std Deviation

The DF test gave a good result this time and we found our value to be 0.00. Let's plot the converted ACF and PACF functions. We can now get an idea of the parameters that will be used to use in ARIMA as seen on figure 17.

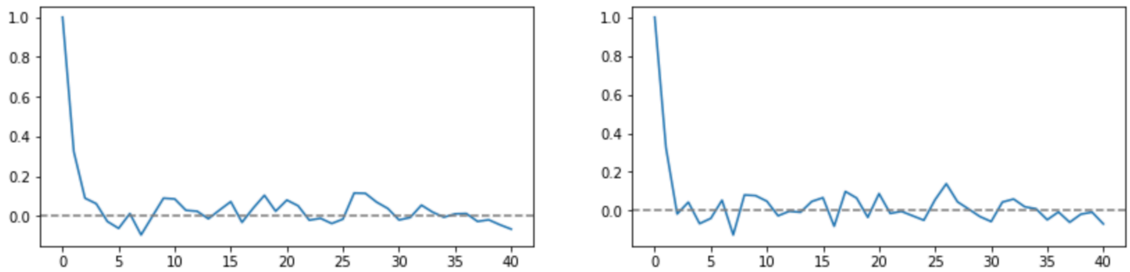

Figure 18: ACF and PACF Functions

We understand from the graph that ACF and PACF approach zero as the delay approaches 1. Assuming d as 1, we will try different p and q values according to the graphs. Next, we'll see the plot of the best model that came out as seen in figure 19.

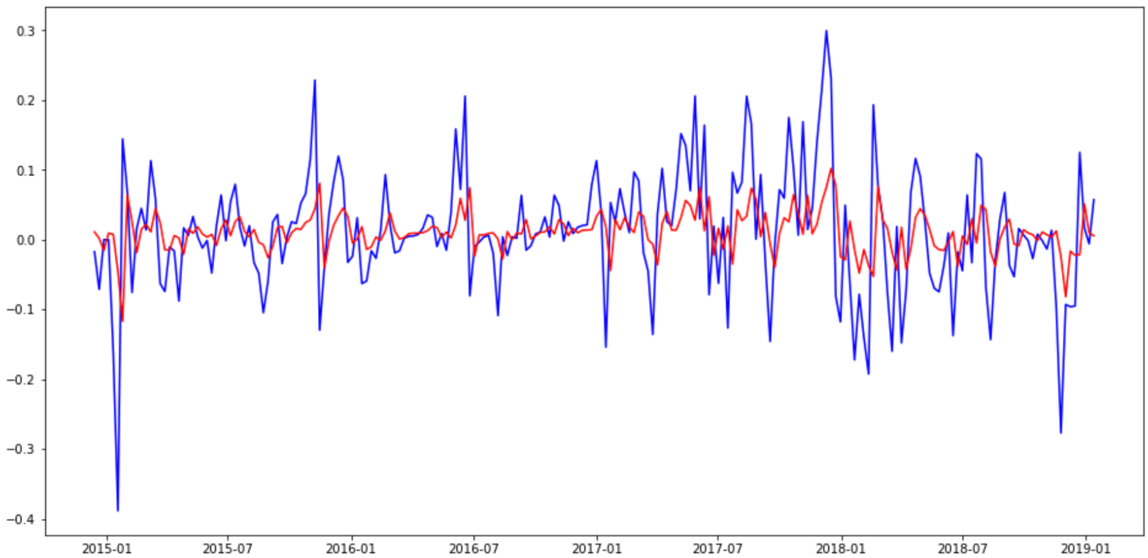

Figure 19: Graph of Best Model (3,1,2)

As shown in the graph above, we got the best result for p 3, d 1, q 2. Now that we know our best model parameters. After the run inverse Box Cox transformation function, we made predictions at ARIMA. The output and plot of the ARIMA table can be viewed in the evaluations section.

4.2. LSTM

For the LSTM Model, keras model (version 2.4.3) was utilized to integrate. Subsequent to stacking the data, the date was changed over to date time utilizing pandas, and the model was discovered utilizing a split_sequence work. Time allotments were appointed and information split into preparing and test sets. We used sequential while building the model. Our target valuable was weighted price. We added 10 embedding layers in LSTM and added none and 1 command to inputshape. After adding 1 Dense layer to the model, we compiled our model after defining mean squared error as loss and determining adam as optimizer. We used 5 epochs and our batch size was 32. A bidirectional model was made for LSTM by using the agent and then the information was prepared. The two-way model was used for measurement and the results were evaluated independently for Bitcoin.

4.3. XGBoost

XGBoost has been implemented with the open-source package keras 2.4.0. The latest version of the package is from github. Supports different weighted characterizations and reviewed target works just as client characterized target usefulness. It is accessible in machine learning dialects and normally incorporates with machine learning pipelines like scikit-learn. The movability of XGBoost makes it usable in numerous environments as opposed to just being attached to a specific stage. Our target valuable was weighted price.

Among strategies, Python's GBM utilizes a voracious methodology that extends just one part of the tree, which makes it quicker, however both scikit-learn and XGBoost give lower precision when learning a complete tree. While both XGBoost and scikit-learn outperform Python's GBM, XGBoost runs multiple times quicker than scikit-learn. In this examination, we found that section subsamples additionally performed somewhat worse than utilizing all highlights. This might be on the grounds that there are key highlights in this informational collection and only one out of every odd element is utilized.

5. EVALUATION

LSTM, XGBoost and ARIMA models were executed effectively and from that point were assessed based on their MAPE, MSE, MAE and R2.

5.1. LSTM Evaluations

R squared presents the outcomes got utilizing LSTM for bitcoin. In terms of R squared, the results provided by the LSTM authorization are far from perfect. The estimated R squared values are about 29.7%, mean square error 246151, mean absolute error is 483.6 and mean absolute percentage error is 8.7%. There are price differences for between recorded price and predicted price.

As can be seen in the graph below, LSTM has done a good job of predicting the trend. Even if the estimation of LSTM is below the price, it seems that situations such as increase and decrease are predicted. Figure 21 shows predicted results for 6 months as below.

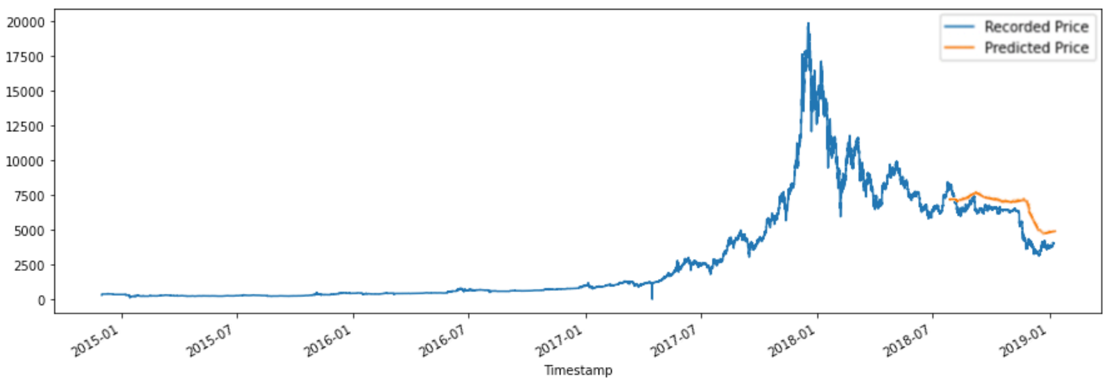

Figure 20: Daily Bitcoin Prices (USD) Predicted vs Actuals for LSTM

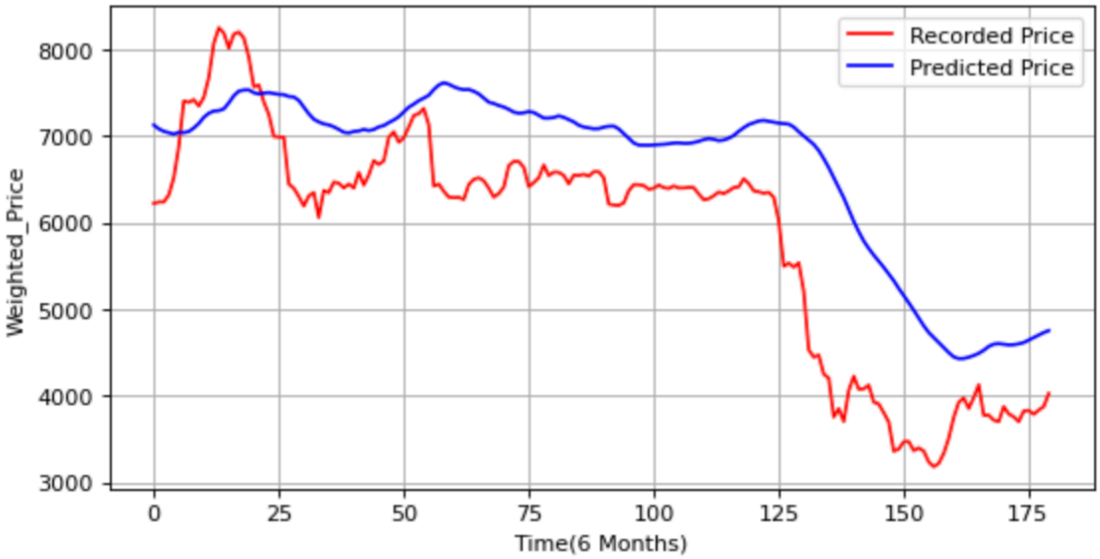

Figure 21: 6 Months Focused Bitcoin price as seen on Recorded (True Result) Prices vs Predicted Prices for LSTM

5.2. ARIMA Evaluations

The standard approach of learning and evaluating time series models is to split the entire time series at a certain time step, where the front part is taken as training the data while the rest is used as testing data. However, such a method could overlook the time-varying behaviour of the volatility series. Consequentially, the derived model has to compromise the non-stationarity in data. Therefore, we adopt a rolling strategy to learn and evaluate models, such that it enables to study the performance of models on different time periods of the data as well as the effect of the look-back time horizon for model learning.

It is critical to expel patterns and seasonalities found in the extricated information, a lot of information was handled before utilizing it as a period arrangement for the ARIMA Model. Subsequent to performing different factual tests, fixed time arrangement was utilized for the Box Cox model and the outcomes acquired had an acceptable proportion as far as likelihood esteems.

Figure 21 shows the price anticipated by the ARIMA and it can plainly seen in bitcoin time series, there is a case of late catching of the trend between recorded and predicted prices. In between 2018 and 2019, the price of bitcoin increased and decreased sharply as shown below Figure 20 and our prediction seen orange line as predicted price.

R squared presents the outcomes got utilizing ARIMA for bitcoin. In terms of R squared, the results provided by the ARIMA authorization are worse than LSTM. The estimated R squared values are about 16.9%, mean square error 545072, mean absolute error is 762.1 and mean absolute percentage error is 11.3%.

Figure 22: Daily Bitcoin Prices (USD) Predicted vs Actuals for ARIMA

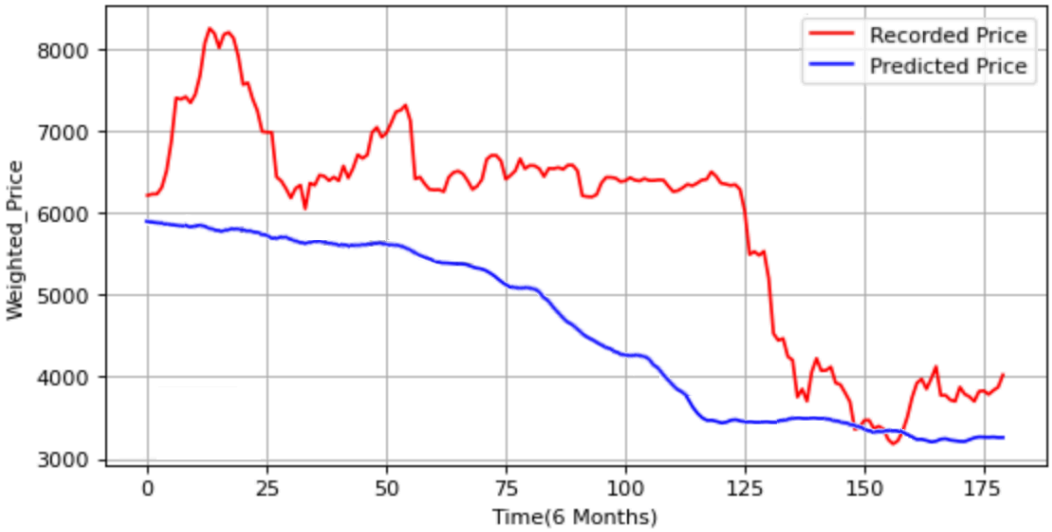

Figure 23: 6 Months Focused Bitcoin price as seen on Recorded (True Result) Prices vs Predicted Prices for ARIMA

5.3. XGBoost Evaluations

Looking at the R square, the XGBoost was the most misleading model and the worst performing model. Despite changing the data, changing the learning rates and changing the max depth, the best result was 18.4% for R squared. Mean square error is 474305, mean absolute error is 396.4 and mean absolute percentage error is 6.1%.

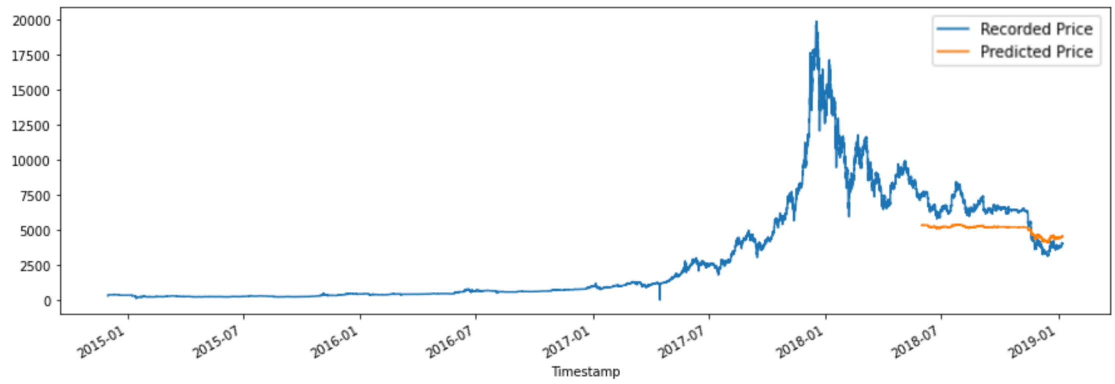

Figure 24: Daily Bitcoin Prices (USD) Predicted vs Actuals for XGBoost

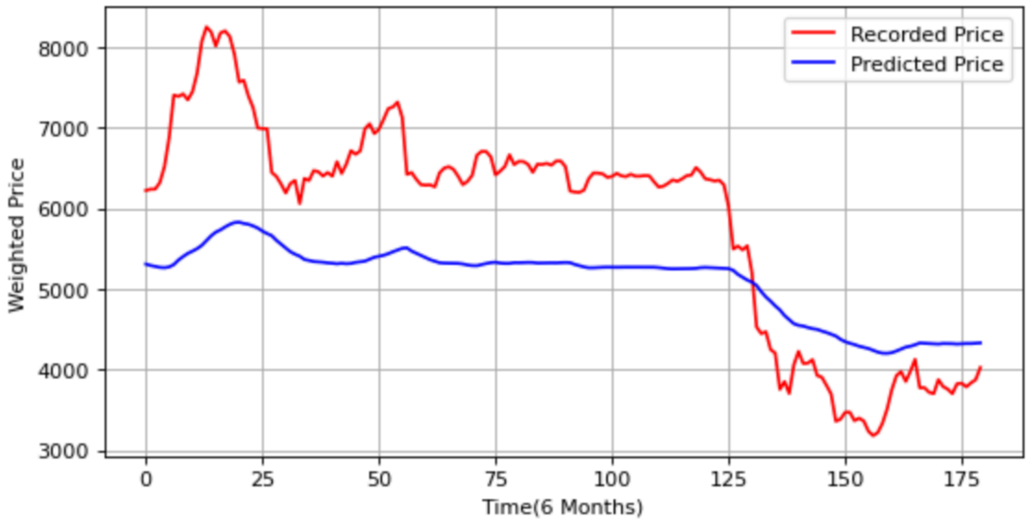

Figure 25: 6 Months Focused Bitcoin price as seen on Recorded (True Result) Prices vs Predicted Prices for XGBoost

5.4. Evaluation Results

Combined dataset consists of 4363457 lines in the bitstamp dataset and 2099760 lines in coinbase dataset. More than 6 million rows in the combined dataset were used in this project. Seasonal patterns of combined dataset are examined via ARIMA and the result shows no existence of strong seasonality. (Wang, 2001)

We used several regression models to predict the 6-month bitcoin price using our feature set and evaluated our performance by mean squared error (MSE). LSTM is currently overestimate, while XGBoost and ARIMA are underestimating. LSTM, XGBoost and ARIMA results can be found in table 1 below. Combined LSTM, XGBoost and ARIMA plots are shown in figure 26.

MODEL	R2	MSE	MAE	MAPE
LSTM	29.7%	246151	483.6	8.7%
XGBoost	18.4%	474305	396.4	6.1%
ARIMA	14.9%	545072	762.1	11.3%

Table 1: Evaluation Results for each model

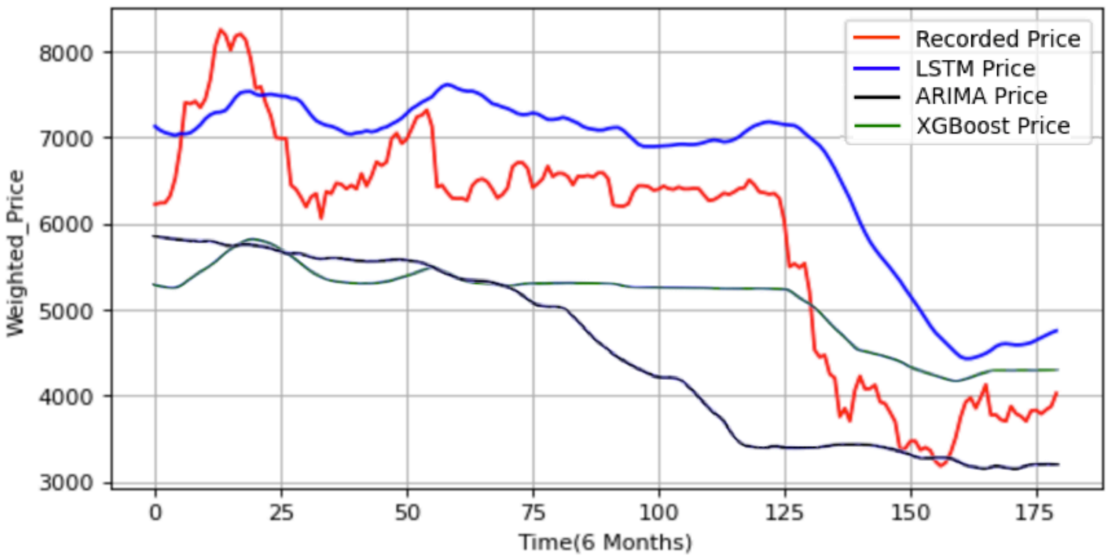

Figure 26: 6 Months Focused Bitcoin price as seen on Recorded (True Result) Prices vs LSTM, ARIMA and XGBoost

As seen in Figure 26, as a result of the operations performed, LSTM is generally estimated to be over. Other methods generally underestimate the same operations. We propose an algorithm by running these three methods for each day and taking the average value. As a result of the average value, we actually do a manual boost. By taking an average of the existing methods named average ML methods, we can find results closer to reality. The results of the average machine learning method we suggested can be found in Table 2 below. The graphic representation of the results is in Figure 27.

MODEL	R2	MSE	MAE	MAPE
Average ML Method	41.6%	288543	420.9	8.1%

Table 2: Evaluation Results for Average ML Methods

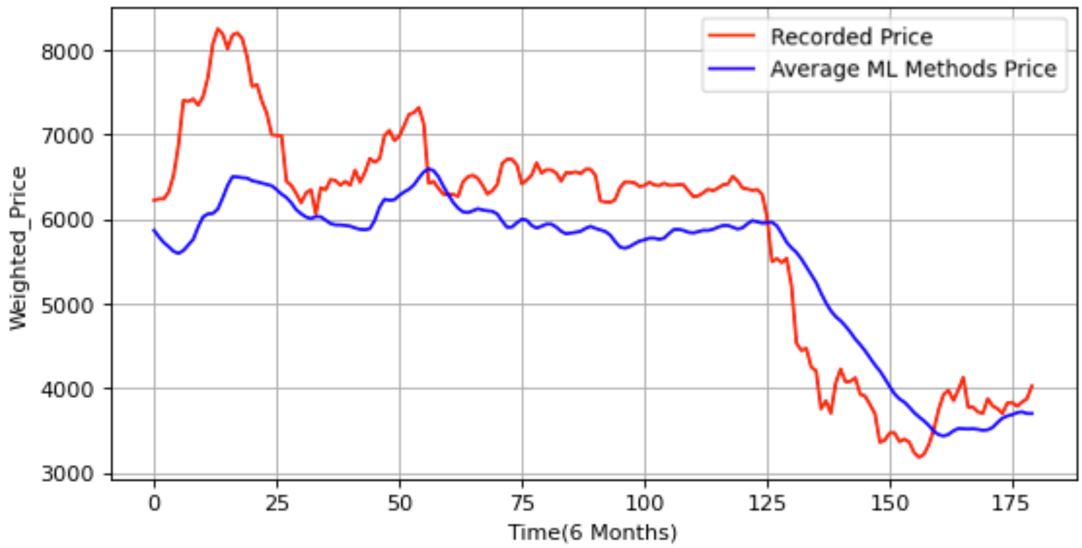

Figure 27: 6 Months Focused Bitcoin price as seen on Recorded (True Result) Prices vs Average ML Methods

As a result of the evaluation, it was determined that the prices were close to each other in terms of value and the data were combined. Price data may be different for each exchange. Every exchange opens or closes a trade on its own prices on their own community. We obtained the daily time series data by determining the Bitcoin weighted price as a dependent variable and Open, Close, High, Low and Volume as independent variable.

6. CONCLUSION AND FUTURE WORK

The experiments in the evaluation section first introduced the expected value results for bitcoin using three different calculations: LSTM deep learning computation, XGBoost scalable tree development model, and ARIMA, a statistics-based factual model. The idea of these three calculations is very unique and each method of use for measuring prices has additionally fluctuated. We predicted the next 6 months with the BARIMA, LSTM and XGBoost methods. The results obtained are summarized in Table 1 as R squares for XGBoost, LSTM and ARIMA. We can say that LSTM surpasses other models among watched and expected prices. In all cases, it should be noted that XGBoost and ARIMA show lower R squared results than LSTM. We compared these predictions using MSE, MAE, MAPE, and R frame performance metrics. The LSTM is the model with the best R squared value of 29.7%.

In addition to the experiments in the evaluation section, it was determined that when the results of three different calculations are combined, we can find a more realistic result on average, and the combination process was performed. At the end of the experiment, the results expressed as the average machine learning method as the R squared value were calculated as 41.6%. It has been observed that different methods give a better result when combined and combined.

Results for LSTM, XGBoost, and ARIMA are normal and not remarkable unlike the top ranked articles, and there could be two possible reasons for this. The primary constraint is that only verifiable bitcoin prices are considered for prediction in this review. Accordingly, it might be wise to include twitter sentimental analysis, dollar index and google trends, which may be responsible for the expansion in digital currency prices. Similarly, high error in general and low r-squared and probability predictions may be due to spikes in digital currency prices in recent years. Because most of the reviews in the literature were using prices until 2017. Major spikes and fluctuations in bitcoin in 2018 are not included in the current articles.

This exploratory study performed a verifiable weighted price-based bitcoin value prediction by combining three machine learning models in addition to using three different machine learning models, and produced remarkable results. The qualities and charts obtained show that LSTM surpasses the ARIMA model like XGBoost when predicting the bitcoin price. When the averages of all the results obtained are taken, the new process appears to be ahead of the LSTM as a forecast. However, these models are not constrained by value expectancy, may perform well for different estimates and assumptions, and can be used in segments other than equity or financial markets. Therefore, each of the three models practically performed well, and it can be said that the neural network works best in this situation.

REFERENCES

Campbell, J.Y., Lo, A.W., MacKinlay, A.C., Whitelaw, R.F. 1996. *The Econometrics of Financial Markets*

Satoshi Nakamoto. (2008). Bitcoin A Peer-to-Peer Electronic Cash System

Christophe Schinckus. (2017). An Essay on Financial Information in the Era of Computerization

Adrian Ţăran - Moroşan, "The relative strength index revisited", African Journal of Business Management Vol. 5(14), pp. 5855-5862, 18 July, 2011

Yulius Hari and Lily Puspa Dewi, "Forecasting System Approach for Stock Trading with Relative Strength Index and Moving Average Indicator", Journal of Telecommunication Electronic and Computer Engineering, Malaysia, Vol. 10 No. 2-3, 2017

Pooja Talreja, "Forecasting Stock Market Trend Using Relative Strength Index", IJLTEMAS, Volume III, Issue VI, June – 2014

Dr. Bhargavi. R, Dr. Srinivas Gumparthi, and Anith. R, "Relative Strength Index for Developing Effective Trading Strategies in Constructing Optimal Portfolio", International Journal of Applied Engineering Research Volume 12, Number 19 (2017) pp. 8926-8936

Sepp Hochreiter and Jurgen Schmidhuber, "Long Short-Term Memory", Neural Computation (1997), Volume – 9, Issue 2, pp. 1735-1780

Chenjie Sang and Massimo Di Perro, "Improving Trading Technical Analysis with Tensorflow Long Short-Term Memory (LSTM) Neural Network", Science Direct, The Journal of Finance and Data Science (2019), Volume 5, pp. 1-11

A. Timmermann and C. W Granger, E_cient market hypothesis and forecasting, International Journal of Forecasting, vol. 20, no., pp. 15- 27, 2004.

Sadegh Bafandeh Imandoust and Mohammad Bolan-draftar. Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock ExchangeS. Bafandeh Imandoust Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 6 (Version 2), June 2014, pp.106-117

Coinbase. 2020. https://developers.coinbase.com/api/v2 [Online]

Bitstamp, 2020. https://www.bitstamp.net/websocket/v2/ [Online]

Bitcoinwiki, 2020. https://en.bitcoinwiki.org/wiki/Bitcoin_history [Online]

Karakoyun, E. and Çıbıkdiken, A. O. (2018). Comparison of arima time series model and lstm deep learning algorithm for bitcoin price forecasting., Proceedings of the Multidisciplinary Academic Conference pp. 171 - 179.

Saxena, A. and 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): 2591 - 2600.

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

XGBoost: Tianqi Chen and Carlos Ernesto Guestrin (2018). A Scalable Tree Boosting System, Carlos Guestrin KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Pages 785–794

Yuqing Dai, Yuning Zhang (2013). Machine Learning in Stock Price Trend Forecasting. Stanford University

I. N. Tansel, S. Y. Yang, G. Venkataraman, A. Sasirathsiri, W. Y. Bao, and N. Mahendrakar, “Modeling time series data by using neural networks and genetic algorithms,” in Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining, and Complex Systems: Proceedings of the Intelligent Engineering Systems Through Artificial Neural Networks, C. H. Dagli, A. L. Buczak, J. Ghosh, M. J. Embrechts, and O. Erosy, Eds., vol. 9, pp. 1055–1060, ASME Press, New York, NY, USA, 1999

C. K. Lee, Y. Sehwan, and J. Jongdae, “Neural network model versus SARIMA model in forecasting Korean stock price index (KOSPI),” Issues in Information System, vol. 8, no. 2, pp. 372–378, 2007

S. Basak, et al., Predicting the direction of stock market prices using tree-based classifiers, North American J. Econ. Finance 47 (2019) 552–567. (Basak, 2019)

L. Wang, K. K. Teo, and Z. Lin, \Predicting time series with wavelet packet neural networks," in Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on, vol. 3. IEEE, 2001, pp. 1593{1597.