Logo, company name

Description automatically generated **Predictive Analytics**

***Causality Analysis between Google trends and the BTC-ETH-ADA Prices***

**Team 2**

*Ashwath Ramesh*

*Kavuri Lakshmi Kausalya*

*Rohit Muralitharan*

**School of Graduate Professional Studies**

Data Analytics

GV – IE 575 - Predictive Analytics

*Spring II Semester, 2022.*

**Document Control:**

**Work Carried out by:**

|  |  |  |
| --- | --- | --- |
| **Name** | **E-Mail** | **Exhaustive List of Tasks** |
| Ashwath Ramesh | akr6021@psu.edu | Data Preprocessing, Data Exploration, Granger Causality Analysis, ARIMA |
| Kavuri Lakshmi Kausalya | lkk5341@psu.edu | Data Preprocessing, Data Exploration, XGBoost |
| Rohit Muralitharan | rxm5874@psu.edu | Data Collection, Data Preprocessing, LSTM |

**Revision Sheet:**

|  |  |
| --- | --- |
| **Date** | **Revision Description** |
| 03-20-2022 | Came up with the project proposal |
| 03-27-2022 | Reviewed the necessary scientific research journal |
| 04-03-2022 | Collected the necessary data |
| 04-10-2022 | Finished with data cleaning and exploration |
| 04-17-2022 | Granger analysis was done, and modelling started |
| 04-24-2022 | Finalized with the model and finished the presentation and report. |
| 04-27-2022 | Project was submitted and presented |

Table of Contents

[**1.** **Introduction** 3](#_Toc101740647)

[**2.** **Problem Statement** 4](#_Toc101740648)

[**3.** **Challenges** 4](#_Toc101740649)

[**4.** **Related Works** 5](#_Toc101740650)

[**5.** **Importance and Impact** 6](#_Toc101740651)

[**6.** **Data Collection** 6](#_Toc101740652)

[**6.1.** **Dataset Description** 10](#_Toc101740653)

[**9.** **Methodology** 14](#_Toc101740654)

[**9.1.** **Granger Causality Analysis** 14](#_Toc101740655)

[**9.2.** **Arima** 14](#_Toc101740656)

[**9.3.** **XGBoost** 15](#_Toc101740657)

[**9.4.** **LSTM** 16](#_Toc101740658)

[**10.** **Results and Interpretation** 16](#_Toc101740659)

[**10.1.** **Granger Causality Analysis** 16](#_Toc101740660)

[**10.1.1.** **Monthly Trend** 16](#_Toc101740661)

[**10.1.2.** **Daily Trend** 17](#_Toc101740662)

[**10.2.** **Arima** 18](#_Toc101740663)

[**10.3.** **XGBoost** 18](#_Toc101740664)

[**10.4.** **LSTM** 21](#_Toc101740665)

[**11.** **Discussion of results** 24](#_Toc101740666)

[**12.** **Feedback** 26](#_Toc101740667)

[**13.** **References** 27](#_Toc101740668)

1. **Introduction**

There's no doubt that the Internet has taken this world by storm and since then, the communication and the significant importance of understanding the difference between real and digital currencies has skyrocketed too. There have definitely been some major hitches and knots like the Dotcom bubble in the late 1990's, but crypto currencies still survived with different currencies reaching peaks at different points in time. And undeniably, there are some companies like Mt. Gox (A Japan company – declared bankruptcy in 2014), Ripple (another cryptocurrency that fell from $3.84 to $0.45 in 2018, giving heavy losses) etc., that faced major losses due to cryptocurrencies. Losses in cryptocurrencies might have many reasons but one of the main reasons is definitely because of the poor prediction, when are prices going to rise, drop and the chances of rising again. Though there are many different digital currencies, in this project, we'll be discussing about BTC (Bitcoin) – ETH(Ethereum) – ADA(Cardano).

So, our project’s main purpose is to predict and analyze the cryptocurrencies (Bitcoin, Ethereum, Cardano) and to find the cause and effect (Causality) between Google Trends and these cryptocurrencies. We want to build a model that predicts the percentage change in each of the cryptocurrencies (BTC-ETH-ADA), with the help of ARIMA (Auto-Regressive Integrated Moving Average), while also considering the factors that affect their values. We also used XGBoost which is an optimized distributed gradient boosting library designed to be highly efficient. It implements machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. Finally we used Bidirectional LSTM to model the data which is a process of making any neural network o have the sequence information in both directions backwards (future to past) or forward (past to future).

1. **Problem Statement**

There are a lot of market investors, who tend to invest in the digital currencies i.e., cryptocurrencies and they tend to look for guides to get them to profit. So, the objective of our project is to design a model for predicting the percentage returns in the cryptocurrencies with the help of ARIMA, XGBoost and LSTM while considering factors that affect their value to rise, drop or rise again or plummet further. We also help the traders in determining whether to take a long or short call for a given trading day.

1. **Challenges**

While the cryptocurrencies are an easy way of making profits but there are many challenges associated with it. One of the challenges faced in the causal analysis of cryptocurrency is that normal causes for changes in market like the country’s GDP, imports and exports and the value of dollar, and the US and world stock market, according to some research papers carried out, does not seem to have any significant impact on the price of cryptocurrency. On the other hand, non-typical factors such as Google trends and social media posts seems to have significant effect. The problem that this pose is that, normally it is comparatively easier to deal with sources like GDP and similar causes since they offer more data and clear insights regarding prices than indirect sources like web searches. This makes it more difficult to build a model that can handle this sort of source.

Another challenge we face is, the cryptocurrency prices tend to be highly volatile. It is difficult to develop an accurate enough model to predict the prices of cryptocurrencies, as they are highly volatile and constantly changing and also have significant fluctuations in exchange rates.

We faced issues in obtaining the data for trend of cryptocurrencies as Google trends only provides data on a month wise granularity. This not only reduces our data points but also predicting for such a long time period is not an effective way based on our studies on cryptocurrency. Thus, with the obtained data scaling techniques were performed on the month wise data to interpolate it into day level granularity.

1. **Related Works**

We can see from the research work carried out by Vasily Derbentsev [1] et.al. that the use of Binary Auto Regressive Tree (BART) has allowed them forecasting the prices of cryptocurrency highly accurately even during dynamic rise and falls in prices.

From the research done by Guandong Xu [2] et.al., we can get an introduction and understanding of causality analysis with fundamental background and key concepts and provide how machine learning can be interpreted using causality analysis.

From the research done by Stefan Haufe [3] et.al., we can see that it helps to estimate the causal interactions in multivariate time series data with the employment of Group Lasso. Complete Granger causality is achieved to find all the casual interactions within the dataset. This is done by finding all the causal relationship in the dataset all at once by using Vector Autoregressive model.

From the research done by [Argimiro Arratia](https://link.springer.com/article/10.1007/s42786-021-00027-4#auth-Argimiro-Arratia) et.al., [Albert X. López-Barrantes](https://link.springer.com/article/10.1007/s42786-021-00027-4#auth-Albert_X_-L_pez_Barrantes) et.al. [4], we can see that Priestley-Subba-Rao a second order stationary test that is used to see how non-constant the time-varying Fourier Spectrum of the series is.

From the research done by [Muhammad Ali Nasir](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8#auth-Muhammad_Ali-Nasir) et.al., [Toan Luu Duc Huynh](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8#auth-Toan_Luu_Duc-Huynh) et.al., [Sang Phu Nguyen](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8#auth-Sang_Phu-Nguyen) et.al., [Duy Duong](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8#auth-Duy-Duong) et.al. [5], we can see that the data series is checked to be stationary using Dickey-Fuller tests. Akaike information criteria is used to choose the lag for the process and Granger causality test to determine the direction of causality. Impulsive response function analysis is performed to have a broader perspective on association among the variables.

From the research done by Yu Huang et.al., Zuntao Fu et.al., Christian L. E. Franzke et.al. [6], we can see that the primary contribution of this research is to disclose how RCC method can be used for the casual inference and to explain the relationship between RCC and Extended Convergent Cross Mapping (ECCM). This research also describes the potentiality of RCC model to identify casual relationships by implementing RCC to time series. This also explains the dominance of RCC in comparison to ECCM.

(Mukherjee et al. 2017) [7] In this paper, the authors aim to discuss about the relationship between the social media conversing and the web searches to understand the impact on the word-of-mouth marketing but in the electronic sense by using the web search engine, Google and uses the 3 major platforms, Instagram, Twitter and Tumblr, for the social media conversing.

(Kristoufek et al. 2013) [8] In this research paper, they are studying the relationship between digital currency and search queries by connecting two of the main phenomena in the recent years – digital currency, Bitcoin and search queries on Google Trends and Wikipedia to show that the search queries and the prices are connected and also to prove that there exists a pronounced asymmetry between them.

(Yang et al. 2015) [9] This research paper puts forth a web search data help to improve the accuracy of forecasting the number of visitors visiting the Hainan Province in China by not only reflecting the trends of people’s preferences but also predicted the future travel behavior while comparing the top two search engines, Google and Baidu

1. **Importance and Impact**

First impression is that there is an interesting relationship between XBTUSD prices and Google Trends for Bitcoin. We see different volatility patterns for prices and Google trends. While price volatility peaks during the period XBTUSD price itself peaked, Goggle trends volatility decreased once prices began to skyrocket, and Bitcoin became a household name. Our aim is to determine if search trend data can be used as a signal to predict Bitcoin prices. Since Google trends data for such a long duration of time is only obtained at a month level granularity, we were able to process and develop a prediction model at a day level granularity. The model is designed to predict the returns a cryptocurrency would generate on a daily basis. Though failing to do so accurately because of reasons like covid impact on the market we were still able to predict a long or short call which will help the traders in making money in the market. We forecast the direction in which the cryptocurrencies price will shoot up thus helping the traders to go for call or put options thus helping the traders earn money when the market goes both ways say when the returns are positive as well as negative.

1. **Data Collection**

The project is focused on mainly the three major cryptocurrencies currently in the market, Bitcoin, Ethereum and Cardano. We have a dataset focusing on the market prices of each of the bitcoin containing time series data from 2013 to 2021. This data was handed to us at the start of the project.

Since the project is focused on the causal links between the market prices and the effect that the search on google has on these, we obtained data from Google Trends for both web and news searches in the United States between 2013 to 2021. The relevant links are provided in the references section. Since we require Google Trends data for a long time frame the data could be obtained only in the units of month rather than the preferred days. So, we performed data transformation to obtain the data at day granularity level. We first obtained the daily data concatenated from multiple 1-month queries and normalized by corresponding weekly trends data. Then we queried for multiple 9-month period with significant overlapping periods and use the overlapped period to have consistent scaling.

Chart, histogram

Description automatically generated

Fig.6.1 Overlapping data

The orange line shows the range for which the weekly data we scrapped is overlapping to making the standardization uniform across the dataset. Then we simply interpolated the daily data from the weekly data.

Chart, histogram

Description automatically generated

Fig.6.2. Daily data

Chart, histogram

Description automatically generated

Fig.6.3. Weekly data

Chart, histogram

Description automatically generated

Fig6.4. Different Methods comparison

Chart, histogram

Description automatically generated

Fig.6.5. Different methods vs original data

We selected the overlapping data since the pattern of the original data was replicated best by this which can be seen in Fig.6.5.

The same method was followed for Bitcoin and Cardano.

* 1. **Dataset Description**

The market dataset consists of 9 attributes and has 2922 data entries for Bitcoin, 2160 for Ethereum and 1374 for Cardano. The attributes are:

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Data Type** | **Example** |
| Name | String (Object) | Bitcoin |
| Symbol | String (Object) | BTC |
| Date | DateTime (Object) | 4/29/2013 23:59 |
| High | Float (Float64) | 147.488 |
| Low | Float (Float64) | 134 |
| Open | Float (Float64) | 134.444 |
| Close | Float (Float64) | 144.54 |
| Volume | Float (Float64) | 0 |
| Marketcap | Float (Float64) | 1603768865 |

Table 6.1. Data description of market price dataset

The Google trend datasets consist of the trends from web searches and News searches obtained from Google trends. The dataset contains information about the number of searches conducted about each of the cryptocurrency and the number of searches has been scaled from 0 to 100, which was observed with a seven-day interval.

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Data Type** | **Example** |
| date | DateTime (Object) | 5/1/2013 |
| Trend | Integer (Int64) | 1 |

. Table 6.2. Data description of Google Trends dataset

1. **Data Preprocessing**

* We checked for any null values in both the trends dataset and the Cryptocurrency dataset and found to have no null values.
* We checked for outliers within the dataset and found that the data didn’t have any outliers.
* The column ‘SNo’ is removed from the cryptocurrency dataset as it is just an identifier and does not have any meaningful insight.
* The ‘Date’ column is transformed to Date-time field.
* We obtained rolling window for 3 and 7 days for both price and trends.
* We also generated lag values up to 3 days.
* We generated a categorical feature for the close value having 0 and 1 with 0 relating to negative returns and 1 relating to positive returns.
* We used MinMaxScaler which scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one.

Chart, histogram

Description automatically generated

Fig.7.1 Normal Distribution of Close Price Before and after Pre-Processing for Bitcoin

Similar distribution to that of one in Fig.7.1 was found in Cardano and Ethereum.

1. **Data Exploration**

* We plotted the daily trends for both the google trends and market prices. From the plots we were able to observe that there was some kind of relationship between the price and trend from the graph plotted.

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Fig. 8.1. Bitcoin Market price Fig. 8.2. Bitcoin Google Trend

We can see from Fig.8.1 and Fig.8.2 that there is some kind of similarity between the trends which should be further explored.

Similar graphs were plotted for Cardano and Ethereum where similar trend was found.

Chart

Description automatically generated Chart, histogram

Description automatically generated

Fig. 8.3. Cardano Market price Fig. 8.4. Cardano Google Trend

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Fig. 8.5. Ethereum Market price Fig. 8.6. Ethereum Google Trend

* We also did the descriptive statistics of the dataset and was able to get a description of the data we were dealing with.
* The data we are handling is a time series data.
* After plotting the correlation matrix, from Fig.8.7 we were able to see that the correlation between the trends and close price for Bitcoin was around 0.71 denoting it has a strong positive correlation. The correlation of trends with other attributes seemed to be the same as with close price, except for volume which reduced to 0.63.

Chart, calendar

Description automatically generated

Fig.8.7. Correlation Heatmap for Bitcoin

Similar and better results were found for Cardano and Ethereum

Calendar

Description automatically generated

Fig.8.8. Correlation Heatmap for Cardano

Calendar

Description automatically generated

Fig.8.9. Correlation Heatmap for Ethereum

* This shows that there might be a causal relationship between trends and the price of the currency which has to be explored.

‘

1. **Methodology**

We first used Granger Causality Analysis to determine the causality effect on month level and day level data between Google trends and Cryptocurrency prices. From this we choose the best granularity level and check their performances with three different models [ARIMA, XGBoost, LSTM]. Since it’s a time series dataset the random split was ignored, and 2021 data is taken as test set and the rest is used for training.

* 1. **Granger Causality Analysis**

Granger causality analysis was undertaken to find the causal link between Trend and the prices of the crypto currencies. F statistics is used for the analysis.

The Null Hypothesis is that trend does not cause changes in prices (p > critical value). The Alternative Hypothesis is that trend does cause changes in prices (p < critical value). The critical value is 0.05.

We did the analysis with both monthly and daily trends.

* 1. **Arima**

We used auto\_arima function to determine the best (p, d, q) values for the ARIMA model and we use it to predict the price returns of the cryptocurrency for the year 2021 and also forecast long and short options based on the increase or decrease in the returns respectively.

* 1. **XGBoost**

We used XGBoost Regressor which is a boosting model to predict the price returns of the cryptocurrency for the year 2021 and XGBoost Classifier to forecast long and short options based on the increase or decrease in the returns respectively.

A picture containing text, person, document, screenshot

Description automatically generated

Fig.9.1. XGBRegressor

A screenshot of a computer

Description automatically generated with low confidence

Fig9.2. XGBClassifier

* 1. **LSTM**

We used Bidirectional LSTM which uses neural network to understand the pattern within the data and remember the previous values during modelling to forecast long and short options based on the increase or decrease in the returns respectively. The activation used was Relu and optimizer used was Adam.

Graphical user interface, text, email

Description automatically generated

Fig.9.3. LSTM Model.

1. **Results and Interpretation**
   1. **Granger Causality Analysis**
      1. **Monthly Trend**

The monthly trends were first used for the causal analysis. The results for each of the crypto currencies are as follows.

Text

Description automatically generated with medium confidence

Fig.10.1. Results of Causal Analysis of Bitcoin for monthly trends

A picture containing table

Description automatically generated

Fig.10.2. Results of Causal Analysis of Cardano for monthly trends

A picture containing diagram

Description automatically generated

Fig.10.3. Results of Causal Analysis of Ethereum for monthly trends

The analysis with monthly trends, we failed to reject the null hypothesis for Bitcoin and Ethereum, which tells us that there is no causal relationship.

We now increase the granularity and take the daily trends.

* + 1. **Daily Trend**

Table

Description automatically generated with low confidence

Fig.10.4. Results of Causal Analysis of Bitcoin for daily trends

A picture containing table

Description automatically generated

Fig.10.5. Results of Causal Analysis of Cardano for daily trends

Table

Description automatically generated

Fig.10.6. Results of Causal Analysis of Ethereum for daily trends

On the other hand, the analysis with daily trends, we rejected the null hypothesis in all three crypto currencies which tells us that there is causal relationship between Google trends and the prices.

With that we decided to proceed with the daily trends data for our modelling.

* 1. **Arima**

We found the best (p,d,q) for close price prediction to be (0,0,0), (0,0,0) , (0,0,0) for Bitcoin, Cardano and Ethereum respectively and for long-short option forecasting (0,0,1), (5,1,0), (1,0,2) respectively.

The RMSE and MAPE values are given in Section 11.

The accuracy for long-short option forecasting was found to be 50.27, 53.53 and 42.78 percentage for Bitcoin, Cardano and Ethereum respectively.

* 1. **XGBoost**

Bitcoin:

Chart

Description automatically generated

Fig.10.1. Feature Importance for Close price prediction

Chart, bar chart

Description automatically generated

Fig.10.2. Feature Importance for Long-Short Option Forecasting

Cardano:

Chart

Description automatically generated

Fig.10.3. Feature Importance for Close price prediction

Chart, bar chart

Description automatically generated

Fig.10.4. Feature Importance for Long-Short Option Forecasting

Ethereum:

![Chart, bar chart

Description automatically generated]()

Fig.10.5. Feature Importance for Close price prediction

Chart, bar chart

Description automatically generated

Fig.10.6. Feature Importance for Long-Short Option Forecasting

The RMSE and MAPE values are given in Section 11.

The accuracy for long-short option forecasting was found to be 48.13, 48.83 and 47.06 percentage for Bitcoin, Cardano and Ethereum respectively.

* 1. **LSTM**

Bitcoin:

Chart, histogram

Description automatically generated

Fig.10.7. Loss and Validation Loss for train and test for Close price prediction

Chart, histogram

Description automatically generated

Fig.10.8. Loss and Validation Loss for train and test for Long-Short Option Forecasting

Cardano:

Chart, histogram

Description automatically generated

Fig.10.9. Loss and Validation Loss for train and test for Close price prediction

Chart, histogram

Description automatically generated

Fig.10.10. Loss and Validation Loss for train and test for Long-Short Option Forecasting

Ethereum:

Chart

Description automatically generated

Fig.10.11. Loss and Validation Loss for train and test for Close price prediction

Chart

Description automatically generated

Fig.10.12. Loss and Validation Loss for train and test for Long-Short Option Forecasting

The Loss and Val Loss for the model was (0.0023,0.0030), (0.0033,0.0041), (0.0036,0.0047) for Close price prediction for Bitcoin, Cardano and Ethereum respectively.

The Loss and Val Loss for the model was (0.0023,0.0030), (0.0023,0.0030), (0.0036,0.0047) for Close price prediction for Bitcoin, Cardano and Ethereum respectively.

The RMSE and MAPE values are given in Section 11.

The accuracy for long-short option forecasting was found to be 48.13, 53.53 and 49.73 percentage for Bitcoin, Cardano and Ethereum respectively.

1. **Discussion of results**

Bitcoin:

Text

Description automatically generated

Fig.11.1 RMSE value for Close Price Prediction

Text

Description automatically generated

Fig.11.2 RMSE value for Close Price Prediction

From the Fig 11.1 and Fig 11.2 we can infer that ARIMA is most suited for Bitcoin

A picture containing antenna

Description automatically generated

Fig.11.3. Model Prediction vs Original Returns

For the Long-Short Option Forecasting ARIMA was found to be the best option based on Results in Section 10.

Cardano:

A picture containing text

Description automatically generated

Fig.11.4 RMSE value for Close Price Prediction

Text

Description automatically generated

Fig.11.5 RMSE value for Close Price Prediction

From the Fig 11.4 and Fig 11.5 we can infer that ARIMA is most suited for Bitcoin

Graphical user interface, chart, line chart

Description automatically generated

Fig.11.6. Model Prediction vs Original Returns

For the Long-Short Option Forecasting ARIMA and LSTM was found to be the best option based on Results in Section 10.

Ethereum:

Text

Description automatically generated

Fig.11.7 RMSE value for Close Price Prediction

Text

Description automatically generated

Fig.11.8 RMSE value for Close Price Prediction

From the Fig 11.7 and Fig 11.8 we can infer that ARIMA is most suited for Bitcoin

A picture containing text, object, antenna

Description automatically generated

Fig.11.9. Model Prediction vs Original Returns

For the Long-Short Option Forecasting LSTM was found to be the best option based on Results in Section 10.

1. **Feedback**

Though failing to accurately predict the price returns because of reasons like covid impact on the market and lacking in perfectly understanding the mood of the public because all those who search about cryptocurrency on google always don’t buy a cryptocurrency and the surge in usage of social media platforms recently all have the effect on our predictions. But the approach of predicting the positive and negative returns and thus recommending long and short options is what makes our model stand out where the traders could earn money on the market going in both directions. From the causality analysis we can see that the lower the granularity level the higher is the cause effect and thus moving forward we plan to work on hour level granularity which will have a better impact in our prediction values. Using data from Twitter and Reddit will also have an significant impact on our model as these data will not only show the trends of the public but also the sentiment of those trends.

1. **References**
2. Forecasting cryptocurrency prices time series using machine learning approach | SHS Web of Conferences (shs-conferences.org)
3. [2006.16789] Causality Learning: A New Perspective for Interpretable Machine Learning (arxiv.org)
4. Challenges in Causality Volume 1 (mlr.press)
5. <https://link.springer.com/article/10.1007/s42786-021-00027-4>
6. <https://jfin-swufe.springeropen.com/articles/10.1186/s40854-018-0119-8>
7. <https://aip.scitation.org/doi/pdf/10.1063/5.0007670>
8. Mukherjee, Partha; Jansen, Bernard J. Conversing and searching: the causal relationship between social media and web search, Internet Research; Bradford Vol. 27, Issue. 5, (2017): 1209-1226. DOI:10.1108/IntR-07-2016-0228.
9. Kristoufek, L. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Sci Rep 3, 3415 (2013).
10. Yang, X., Yang, X., Pan, B., Evans, J.A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. Tourism Management, 46, 386-397.