DATA301 GDELT Project

George Carr-Smith
69583529

6/4/2021

1 Abstract

As the price of bitcoin reaches new peaks, its prevalence in the worldwide news increases as media outlets become more vocal on the subject. The project takes the daily volume and tone of the mentions of “bitcoin” in online news, and correlates them with the price of bitcoin in order to find correlation and predictive qualities between the datasets. The volume timeline and tone timeline of bitcoin mentions are taken from the GDELT Online News Summary and the daily price of bitcoin is taken from Coindesk’s historical bitcoin price data. The cosine distance algorithm is used to find correlations between the data sets and multiple linear regression is used to find the predictive qualities between the data sets. The intended result is to find correlation and predictive qualities between the mentions of bitcoin, their tone and the price of bitcoin in order to predict the price of bitcoin.

2 Introduction

2.1 Background

The data that is used in the project is from the GDELT Online Summary API which is a summary of the global news related to a query keyword. In this case, the query keyword was bitcoin. The data is restricted to what online news the GDELT project monitors. From the summary generated, data from the Volume Timeline and Tone Timeline sections were used. The Volume Timeline data contains volume intensity which is the percent of monitored online news published that day that contains the keyword bitcoin. The Tone Timeline contains the average tone of the monitored online news published that day that contains the keyword bitcoin. The data of the prices of bitcoin contains the fields Date, Opening Price and Closing Price. The Opening price and Closing Price fields are the price of bitcoin at exactly 12:00 AM when the day starts and when the day ends.

The cosine distance algorithm is a distance measure which measures the similarity between two vectors. The program will use the columns of data points as vectors and the algorithm will conclude if the columns are similar. The linear and multiple regression algorithms model the relationship between dependent variables and an independent variable. The regression algorithms conclude whether there is a predictive relationship between the variables.

2.2 Research Question

What is the relationship between the price of bitcoin and the number of mentions of bitcoin in the global news and the tone of the articles that mention it?

2.3 Relevance of Research Question to Data Set

The research question is relevant to the dataset as the number of mentions of bitcoin can be found in the Volume Timeline of the GDELT Summary and the tone of the articles can be found using the Tone Timeline of the GDELT Summary. Implementing the cosine distance algorithm will find the similarity relationship between the price dataset and the volume and tone datasets. The multiple linear regression algorithm will look for a predictive relationship between the datasets.

3 Experimental Design and Methods

3.1 Algorithm, Data Flow, and Program Design

The csv files containing the tone timeline and the volume timeline are downloaded from the GDELT Summary api using the urllib.request library. The bitcoin price is downloaded from the coindesk website using the same library. Once the files are saved within the colab file browser, they are turned into pandas dataframes using the .read_csv function which reads a csv file into a pandas dataframe. Once that is completed, there are the dataframes containing bitcoin daily opening and closing prices, daily average tone values and daily volume intensity values with their respective dates. The program then calculates the average price of bitcoin for the day by taking the mean of the opening and closing prices for the day. Using the new average price column, a price change column is calculated by taking the average price for each day and subtracting the price of the previous day. The columns containing the values of volume intensity, average tone, bitcoin average price, bitcoin price change and their respective dates are copied into a new dataframe called the processing dataframe which will be used by the cosine similarity and linear regression algorithms. The processing dataframe was created so the data is lined up correctly with its dates so the data processing algorithms do not need to compare the dates of each data point before processing them. Pandas dataframe transformations are then used to ensure the values in each row correspond to their respective dates in the original data files. The dataframe is then filtered so it only contains rows with dates between the startdate and enddate variables.

The cosine distance algorithm finds the similarity between two columns which are converted to vectors by finding the angle between them. The cosine distance algorithm finds the angle between two vectors by dividing the dot product of the vectors by the product of their magnitudes and subtracting that from one. The program uses the cosine function in the scipy spatial distance library to do this calculation.

The linear regression algorithm tries to find a linear equation which suits the relationship between bitcoin price, the volume of mentions of bitcoin and the average tone of the mentions. The program uses the cuML library from NVIDIA RAPIDS to model the linear regression using a cuDF dataframe as the dependent variables and a single column of a cuDF dataframe as the dependent variable . The processing dataframe must be converted from a pandas dataframe into a cuDF dataframe. The dataframe is converted by using the pandas to_csv() function to turn the dataframe into a csv file, the csv file is then read into a cuDF dataframe using the cuDF.read_csv() function. The bitcoin price column of the cuDF dataframe is set as the independent variable while the mentions and tone columns in the dataframe are set as the dependent variables. The cuML train_test_split() and predict() functions are then used to model the multiple regression.

3.2 Description of Specific Code and Libraries

- Pyspark library: used for RDD’s
- Urllib.request library: downloads the data from their respective URLs using the downloadFile() function
- Pandas library: used to pull the data from the downloaded files into dataframes which can be processed by the cosine distance algorithm
- Cosine function from the scipy.spatial.distance library: used to calculate the cosine distance between columns of the pandas dataframes
- cuDF library: allows the data to be put into a GPU dataframe for use by the multiple regression algorithm
- cuML library: implements the multiple regression algorithm.
- Train_test_split function from cuML: used to evaluate the multiple regression algorithm
- mean_squared_error function from cuML: finds the mean squared deviation between the estimated and actual values
- Matplotlib.pyplot: used to create graphs

4 Results

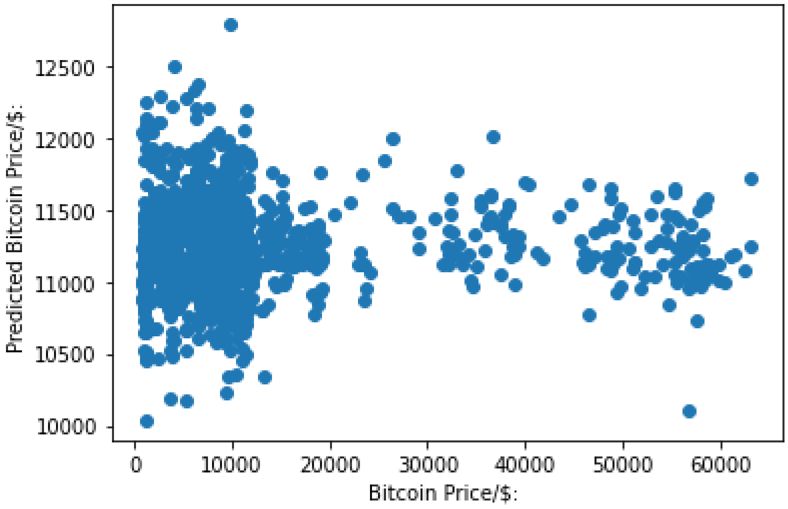
Cosine Distance Algorithm:

1. The cosine distance between Bitcoin Price and Mentions is 0.3106494267982728
2. The cosine distance between Bitcoin Price and Tone is 1.2467356313990057

Multiple Linear Regression:

- Linear Regression with Volume intensity of mentions and average tone as the independent variables and bitcoin price as the dependent variable

Graph of the Multiple Linear Regression Prediction vs Actual Values



:

```
Average difference between the predicted value and the actual
value = 11768.84429004218
```

4.1 Results Drawn from Data

The cosine distance algorithm calculated that the cosine distance between the vector of bitcoin prices and the vector of volume intensity of bitcoin mentions were similar as they had a cosine distance of 0.31. This signifies a correlation as there is a small angle between their vectors. The cosine distance algorithm calculated that the cosine distance between the vector of bitcoin prices and the vector of average tone of bitcoin mentions were dissimilar as they had a cosine distance of 1.24.

The linear regression algorithm could not accurately estimate how the price of bitcoin changes as the volume intensity of the mentions of bitcoin and the average tone changed. When using the volume intensity and average tone as the independent variables and bitcoin price as the dependent variable, the program created a multiple linear regression model which predicted the price of bitcoin. When plotting the predicted price against the actual price, the graph did not form a straight line. A straight line would signify that the dependent variables could be used to accurately predict the independent variable.This concludes that there is not a predictive relationship between bitcoin price, volume intensity and average tone relating to bitcoin.

5 Conclusion

I was not able to completely answer my hypothesis question as the multiple linear regression did not find a predictive relationship between the price of bitcoin, volume of its mentions and average tone. The multiple linear regression model could not accurately predict the price of bitcoin for a few possible reasons. The first reason is that there is not enough data available to make an accurate prediction as the GDELT Summary data used only goes back until 2017. Another reason could be that there are other independent variables that I had not considered which could more accurately estimate the price of bitcoin.

One implication from the results is that the graph of bitcoin price will be similar to the graph of the volume intensity of the mentions of bitcoin. Another implication is that global news most likely does not affect the price of bitcoin.

If continued in the future, I would look into other sets of data and their relationships with the price of bitcoin. Examples of other data sets that would prove useful are: stock market data, bitcoin mentions on the internet, bitcoin mentions by Elon Musk and mentions of bitcoin by the Chinese government. I would continue creating multiple linear regression models with these data sets in order to try to find data sets which can accurately predict the price of bitcoin. A future question I now have is if it is possible to accurately predict the price of bitcoin.

6 Critique of Design and Project

One part of my design that could have worked better with a different approach was finding the correlation between the data sets by using cosine distance. Cosine distance is usually used as a distance measure for comparing the similarity of text documents or images so it does not provide meaningful information on the relationship between the price of bitcoin and the volume intensity of its mentions.

To fix the design flaw, I could have used a different algorithm to find similarities or I could have delved deeper into machine learning algorithms which could find correlations between the data sets.

7 Reflection

7.1 Course Concepts and Tools Useful to Project

- Cosine distance formula
- Cuda model
- Resilient Distributed Datasets labs and lectures
- Linear regression colab demonstration notebook and lab 8

7.2 Concepts Learned

While completing the topic I learned how to use new concepts and libraries related to data processing. I learned what a dataframe is and how to use the pandas library to create, edit and perform functions on dataframes. I also learned how to use the cuml library and cudf dataframes to perform linear regression on sets of data. The project taught me how to extract and interpret data from The GDELT Project. While implementing linear regression to my project I learned the steps involved in modelling it. The most important learning experience I gained from this project was how to read and understand documentation in order to implement new libraries.

8 References

- Provide any citations and/or links to notebooks, datasets, etc

Notebooks used:
https://colab.research.google.com/drive/1hXAeG6yheFUQiHfc9Z5ISfNBqAQw47Dq
https://colab.research.google.com/drive/1LE8BEG-4k5m5-i9iK8W4FAuK8wg-IZS4
https://colab.research.google.com/drive/1sTsI_-f2ipgzqM6htsVdKjf4MZ3Ds2CW#scrollTo=fVf4e7R5gM-n
Citations:

[1]The GDELT Project. (2017–2021, January 1–June 2). GDELT Online News Summary [Dataset]. https://api.gdeltproject.org/api/v2/summary/summary

[2]Bitcoin Price Index — CoinDesk 20. (2017–2021, January 1–June 2). [Dataset]. https://www.coindesk.com/price/bitcoin

[3]Robson, W. (2020, August 31). Beginner's Guide to Linear Regression with cuML - Future Vision. Medium. https://medium.com/future-vision/beginners-guide-to-linear-regression-in-python-with-cuml-30e2709c761

[4]Calculating cosine similarity across column in pandas. (2016, June 20). Stack Overflow. https://stackoverflow.com/questions/37921237/calculating-cosine-similarity-across-column-in-pandas