Cryptocurrency Jumps using Reddit data

Arthur MARON

April 14, 2023

Abstract

This project aimed to predict daily jumps in the price of Bitcoin using Reddit data. The project involved collecting data on the number of submissions and comments related to Bitcoin topics on Reddit, and performing sentiment analysis on the text data using natural language processing (NLP) techniques. The sentiment analysis results were used to create features for different predictive models, which were trained using supervised learning techniques. Models were evaluated on a test set to assess their accuracy in predicting daily jumps in the price of Bitcoin. However, further testing and research are needed to improve the accuracy of the models and to explore other potential sources of data for predicting Bitcoin jumps.

1 Introduction

1.1 Description of the project

The main objective of this project is to use data obtained from the Reddit API to anticipate daily unusual fluctuations in the value of Bitcoin. To accomplish this, we will gather data on the quality and quantity of Bitcoin-related submissions and comments from Reddit, and the text data will be analyzed using NLP methods to determine sentiment from reddit users. This information will be utilized to build predictive models that can categorize whether the price of Bitcoin will significantly rise or fall on a given day. The model will be trained using historical data and tested to measure its precision.

1.2 Project workflow

The project workflow can be divided into the following steps:

1. **Data Collection:** The first step will be to collect data on Bitcoin-related submissions and comments on different subreddits. This can be done using the Reddit API, which allows for the retrieval of posts and comments based on specific search terms. The data will be temporary stored in txt files to be pre-processed.

2. **Data Preprocessing:** Once the data has been collected, it will need to be cleaned and pre-processed. This will involve removing any irrelevant or duplicate data, taking the daily difference, normalization, and other techniques to prepare the data for our model. Our goal is to obtain a clean and simple dataframe.

3. **Sentiment Analysis:** The reddit data will be analyzed using NLP techniques to perform sentiment analysis on the text data. This will involve identifying the sentiment of each submission or comment as positive, negative, or neutral, and identifying certain key terms specific to internet forums that could translate a certain market sentiment.

4. **Model Training:** The features extracted for NLP analysed and the market data from cryptocurrencies will be used to train a predictive model using supervised learning techniques such as logistic regression, decision trees, random forests... Different models will be implemented to compare results, and a final model aggregating all models will also be studied.

5. **Model Evaluation:** The model will be evaluated on a test set to assess its accuracy in predicting daily jumps in the price of Bitcoin. The test set will be separate from training data, as not to inlcude future information in the training of the algorithm.

1.3 Tools and Libraries

This project will be implemented in Python for simplicity, and we will need to use various libraries such as *pandas*, *numpy*, *matplotlib*, *cryptocmd* for scrapping cryptocurrencies, *request* for scrapping Reddit data, and other libraries for classification models.

2 Data Collection

This project requires two different types of data : bitcoin prices and data from crypto-related subreddits. We will utilize the *request* library to build a scrapper for subreddit, and use the existing scrapper for crypto.

2.1 Data selection

2.1.1 Reddit data

The first step in data collection is identifying the source we will scrape reddit data from. There are hundreds if not thousands of subreddits related to cryptocurrencies, but we choose to focus on those who have the most traffic and are the most likely to be commented on by internet traders and speculators. We identified the following subreddits as candidates:

1. **r/CryptocurrencyMemes** The CryptocurrencyMemes subreddit is a community on the social media platform Reddit that is dedicated to sharing memes and humor related to cryptocurrencies. The subreddit has since grown to become one of the largest and most active communities dedicated to cryptocurrency humor. Members post a wide variety of content, including memes that poke fun at the volatility of crypto prices, the trends surrounding new coins, the jargon and technical language used in the crypto world, and the sometimes bizarre behavior of cryptocurrency enthusiasts.

2. **r/Wallstreetbets** Amateur traders and retail investors discuss and share investment ideas, particularly in the context of buying and selling options and stocks. In January 2021, the WSB community gained mainstream attention when it orchestrated a coordinated buying campaign of shares in GameStop (GME), a struggling video game retailer. The idea was to buy up shares of GME, driving up its price and causing a "short squeeze" for the hedge funds that had bet against the company by shorting its stock. In fact, it is this story that inspired this project. While the GME saga has died down, the WSB community remains active and continues to influence investment trends and market sentiment.

3. **r/Cryptocurrency** This subreddit serves as a hub for crypto enthusiasts, investors, traders, developers, and anyone interested in learning about the latest trends and developments in the world of cryptocurrencies. Members of the Cryptocurrency subreddit post and comment on a wide range of topics related to cryptocurrencies, including news and analysis of market trends, discussions about the technology behind different coins, debates about the pros and cons of different investment strategies, and advice for newcomers who are just starting to explore the world of crypto.

2.1.2 Cryptocurrency data

We will focus our attention on Bitcoin as it is the most famous and go-to cryptocurrency for beginners and reddit investors. Our models are able to take several cryptos into account, but for simplicity, only bitcoin jumps we will be tracked in the beginning.

2.2 Data extraction

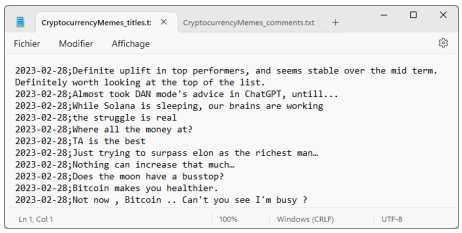
Our code is a Python script that scrapes Reddit comments and submissions from several subreddits, and retrieves data on cryptocurrency prices, comment counts, and scores, and then merges all the data into a single dataframe that is saved as a CSV file.

2.2.1 Reddit data

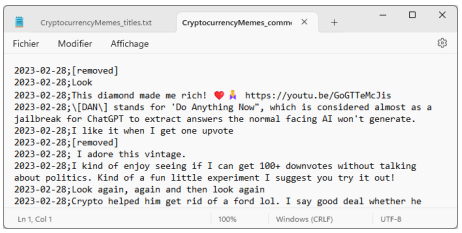
Our code is using the Pushshift API to download comments and submissions from a specific subreddit. Pushshift is an online data collection and analysis platform that allows you to search and retrieve data from various sources on the internet. To achieve this, the code will ask for two

dates - a start date and an end date. Between these two dates, the code will retrieve all submissions and comments that were made in the specified subreddit and store them in a txt file.

Figure 1: Examples of subreddit data



(a) txt file with titles



(b) txt file with comments

2.2.2 Cryptocurrency data

Next, our code gets data for cryptocurrencies using the CoinMarketCap API. Then, for each cryptocurrency we are interested in, the code creates a CmcScraper object with the given start and end times, and retrieves the price data as a DataFrame.

3 Sentiment Analysis and data preprocessing

3.1 Reddit data (explanatory data)

Regarding the reddit data, we first use the **SentimentIntensityAnalyzer** class to perform sentiment analysis on extracted titles and comments. This method returns a compound score which is a value between -1 and 1 that represents the overall sentiment of the text. A score closer to -1 indicates negative sentiment, a score closer to 1 indicates positive sentiment, and a score near 0 indicates neutral sentiment.

On top of that, we will check for certain keywords from a corpus of words we chose ourselves that translate sentiment of buy/short market orders form reddit users. We will compute a **corpus** score for the frequency of appearance of these keywords for each title and comment between -1 and 1. A score closer to -1 indicates a sell sentiment, a score closer to 1 indicates a buy sentiment, and a score near 0 indicates neutral sentiment.

Finally, we also monitor the volume of submissions and comments made on each day, to have a feature dedicated to internet traffic on subreddits.

The data is aggregated to have a mean score for each day, and will therefore use averaged data for each day of the training set. We will also apply daily difference for the reddit data to utilize the dynamics of NLP scores, as standalone scores won't be of much significance. Every column will be normalized with the min-max method a data set appropriated for machine learning algorithms. In the end, we have converted our collection of titles and comments into a dataframe containing three features for each subreddit: **compound score**, **corpus score**, and **count**. Our goal is now to use these features to predict a cryptocurrency jump.

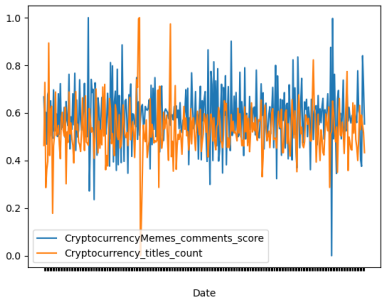
3.2 Crypto data (explained data)

Regarding our cryptocurrencies, we are interested in whether there are daily jumps. Therefore, we determine if a cryptocurrency has jumped if the daily variation is significant (higher than a specified daily threshold). To do this, we compute the daily maximum variation and normalize it by the average price of the day. This gives us a ratio that we convert into 1 if it surpasses the threshold, or 0 it is doesn't. We therefore obtain a dataframe of zeros and ones, ready to use for our logistic regression modelling, or classification models.

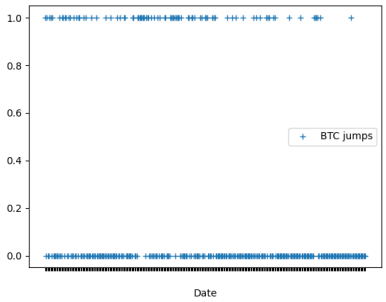
The threshold we use has been determined empirically, by trying different values and checking market data to confirm if detected jumps indeed occurred in real life. This parameter is very important as it changes the output of the target variable, and should be adjusted carefully according to the jump-sensitivity we wish to detect.

In the end, we have normalized daily changes of NLP scores of three subreddits (titles and comments) and jump-labeled data for bitcoin. Over a year, we therefore have 364 days, and 15 features in our explanatory data.

Figure 2: Examples of features in data after preprocessing



(a) Reddit data



(b) Bitcoin data

4 Model Training

4.1 The different models used

4.1.1 Logistic Neural Network

The model is designed to predict whether or not there will be a daily jump in the price of Bitcoin based on input data, which represents the NLP analysis of submissions and comments on chosen subreddits, and output data, which represents the target variable (1 for a price jump, 0 for no jump). The model consists of a series of linear layers followed by **ReLU** activation functions, which are used to introduce non-linearity into the model. The final output layer uses a **Sigmoid** activation function to output a value between 0 and 1, which represents the model's confidence in its prediction. The model is trained using the back-propagation algorithm with the Adam optimizer, and the number of epochs, batch size, and learning rate can be adjusted to optimize performance. The output of the model is a probability value, which is our prediction of whether a daily jump in the price of Bitcoin is likely or not.

4.1.2 Feed Forward Neural Network

The model is designed to predict our un-categorized Y data, which is the maximum variation in the day. The model consists of a series of linear layers followed by **ReLU** activation functions, which are used to introduce non-linearity into the model. As before, the model is trained using the back-propagation algorithm, epochs, batch size, and learning rate. The output of the model is a value which is our prediction of the variation of bitcoin this day. We can then decide whether it is a jump if it is greater than a certain threshold.

4.1.3 Logistic Regression

Logistic regression is used to predict binary outcomes, where the target variable can take only two possible values, such as 0 and 1. The algorithm will learn a set of coefficients for each input feature that best separate the classes of whether the daily price of Bitcoin increased or decreased significantly. The logistic regression model uses the sigmoid function to map the linear combination of input features and coefficients to a probability of the target variable being 1 (jump in Bitcoin price). Once the logistic regression model is trained, it can be used to predict whether the daily price of Bitcoin will jump or not based on the values of the input features.

4.1.4 XGB Classifier

A XGB Classifier is a type of gradient boosting machine learning model that uses an optimized implementation of the gradient boosting algorithm called XGBoost. It is commonly used for classification tasks where the goal is to predict the class or category of a given input data point. The XGB Classifier works by iteratively adding decision trees to the model, where each subsequent tree attempts to correct the errors of the previous tree. During training, the model learns to assign higher weights to the data points that are more difficult to classify correctly. The XGB Classifier is a powerful and flexible model that can achieve high accuracy on a wide range of classification tasks, especially when the data is structured and well-suited for decision tree-based models.

4.1.5 Isolation Forest classifier

The Isolation Forest classifier is an unsupervised machine learning algorithm that is used for anomaly detection tasks. It is based on the concept of decision trees and operates by isolating anomalous data points using a series of randomly constructed decision trees. The algorithm works by randomly selecting a subset of features and partitioning the data points based on their values along these features. The process is repeated recursively until each data point is isolated in its own partition or until a predefined depth of the tree is reached. The number of partitions required to isolate a data point can be used as a measure of its anomaly score. The Isolation Forest classifier has a number of advantages, including its ability to handle high-dimensional data sets and its insensitivity to the scale of the data. It also works well with both small and large data sets.

4.1.6 Tree classifier

A tree classifier predicts the value of a target variable based on a set of input features. The algorithm will first split the input data based on the feature that best separates the classes (0 and 1). Then, the process of splitting and classification will continue recursively for each branch until the tree is fully grown or some stopping criterion is met. Once the decision tree is trained, it can be used to predict whether the daily price of Bitcoin will jump or not based on the values of the input features.

4.1.7 Random Forest classifier

Random forest build multiple decision trees and combines their outputs to make predictions. The algorithm will train multiple decision trees on different subsets of the input data, each tree using a random subset of the features (number of submissions, comments, sentiment scores) to make its decision. The output of each tree will then be combined to make a final prediction using a majority vote or weighted average.

The random forest model is more effective as it is able to handle high-dimensional data and capture non-linear relationships. Additionally, it can help to reduce overfitting by combining the outputs of multiple trees. By optimizing the hyperparameters of the model, such as the number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to split a node, we can improve the accuracy of the model and predict daily jumps in the price of Bitcoin with a high degree of confidence.

4.2 Training set

To test the accuracy and effectiveness of the predictive model built in this project, data from the whole year 2022 will be used. The model will be fed with daily data from reddit and crypto markets as input, and the predictions made by the model will be compared to the actual price jumps of Bitcoin during that year. This comparison will enable us to evaluate the accuracy and reliability of the model, and to identify any areas where improvements can be made.

5 Model Evaluation

5.1 Testing set

The evaluation of the predictive model will be carried out on the first three months of 2023. The model will be fed with data from this period, consisting of the number of submissions and comments related to Bitcoin on Reddit, along with their corresponding sentiment scores. The model will then generate predictions for each day in this period. These predictions will be compared to the actual price jumps of Bitcoin during the same period to determine the accuracy of the model. By testing the model on this data, we can determine whether it is reliable and accurate in predicting daily jumps in the price of Bitcoin.

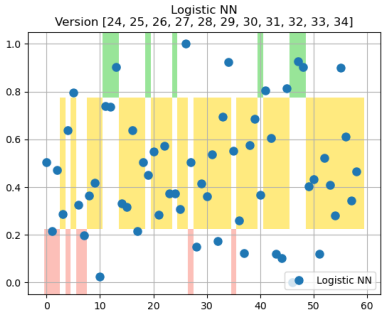
5.2 Testing results

The results vary from model to model, and are not as precise as we expected. We find that some models are close to predicting a jump on the right day, others identify a jump in the surrounding days, and others miss completely. We have the following accuracy ratio of classification for every model :

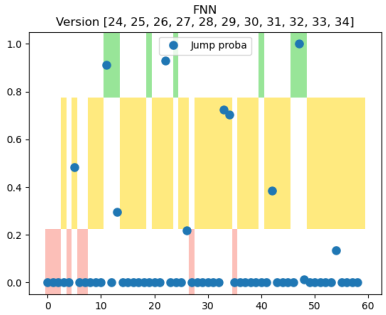
Logistic neural network	0.729
FNN	0.898
Logistic regression	0.898
XGB classifier	0.881
Isolation Forest	0.797
Tree	0.763
Random Forest	0.881

What we find is that Logistic regression and XGB have the best scores overall. This would need to be confirmed with a larger dataset for training and testing.

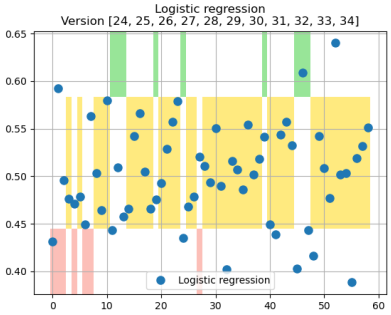
Figure 3: Results of the 7 models



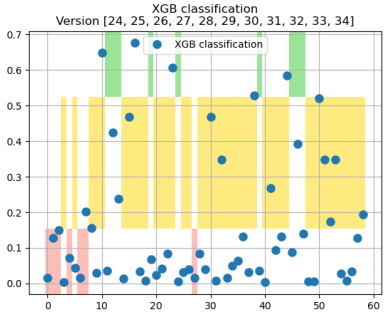
(a) Logistic Neural Network



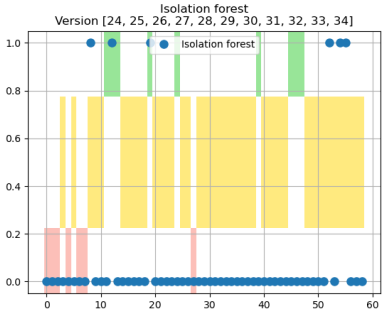
(b) Feed Forward



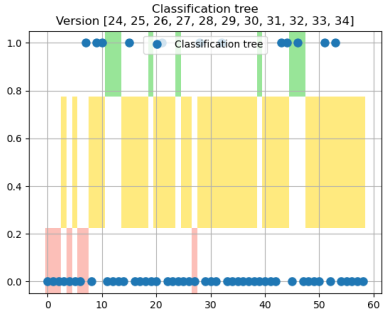
(c) Logistic regression



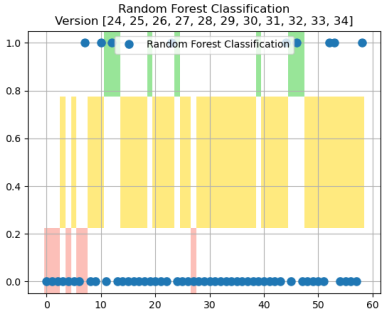
(d) XGB classificator



(e) Isolation forest



(f) Classification tree



(g) Random forest

To see if aggregating model results would have any effect on accuracy, we decided to analyze and group the different results from models according to two methods. The first one is by taking the mean of each model, and the second is taking the label that was the most predicted by the models. We have the following results :

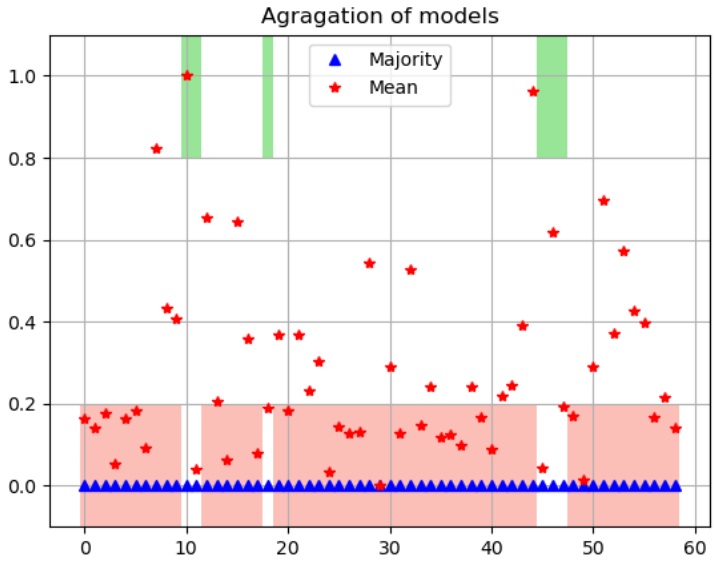


Figure 4: Results of different models on evaluating data set

6 Conclusion

While the project showed promising results, further testing and research are needed to improve the accuracy of the models and to explore other potential sources of data for predicting Bitcoin jumps. Nonetheless, this project highlights the potential of using social media data and NLP techniques for predicting cryptocurrency price movements, and could lead to more effective investment strategies and risk management in the future.

7 References

1. **Github access :** link
2. **Subreddits choice :** link
3. **News article about reddit's influence over trading :** link
4. **Game Stop story :** link