Real-time Cryptocurrency Prediction and Analysis Platform

Chengxi Li, Haopeng Wang, Michael Yang, Hao Zheng
MSc. in Computing Science, Big Data
Simon Fraser University

1 Motivation

After the revolutionary introduction of the world first decentralized cryptocurrency, bitcoin in 2009, the popularity have grown rather slowly due to its unconventional idea of 'proof-of-work' and complex nature of blockchain technology.

However, only after a few years of hibernation, cryptocurrency and blockchain has become the buzzwords in the new world financial market. There are so many investors out there looking for coin information and investment guides to help them make a profit in this financial blue ocean.

By discovering this demand, we are trying to build a one-stop web application for cryptocurrency enthusiasts. However, none of the existing cryptocurrency website is capable of actually giving investment suggestion and price predictions. So in this project, we aimed to break this limitation by two steps.

Firstly, we decided to build a website with comprehensive trading and social information about cryptocurrency market making sure that our users are fully aware of the current trend or any potential investment opportunity.

Secondly, we will provide the users with real time coin returns' prediction by empowering our website with a deep learning model so that the users can use this prediction as a guide to help them making smarter investment decisions.

2 Problem Statement

2.1 Goals

Our primary goal is to build a Realtime bitcoin price prediction streaming system that can process the second-level price related data and predict the price in the next second. During the project, we wish to answer the following questions:

What are the factors that have the biggest impact on cryptocurrency price ?

In this report, we used visualization tools such as seaborn and matplotlib to help us visualizing the various analysis we employed to our data. For example, we used statistical methods such as correlation, time-series analysis and etc to find out the features that are likely to affect the price of cryptocurrency. We also conduct sentiment analysis on cryptocurrency-related news to generate numerical features, and get social data through API such as twitter page behaviors, reddit comments, number of market and analysis views, and etc, to see if such features would affect the price of cryptocurrencies.

What would be the price of cryptocurrency in next minute/hour/day ?

Based on historical minute data, sentiment data and other related features mentioned above, we build a pipeline to predict the minute-level prices of cryptocurrencies in real time. We use a three-layer bidirectional LSTM model for price prediction.

Can we visualize the cryptocurrency ecosystem through a one-stop web frontend ?

We build a one-stop web frontend to visualize and analyze the openly available cryptocurrency data, and draw some informative and dynamically updated graphs to present useful information for cryptocurrency enthusiasts. Our web application is able to provide dynamic prediction result to guide investors with their decisions.

2.2 Challenges

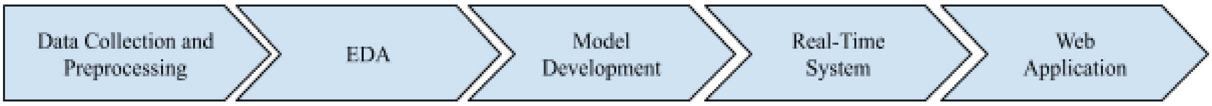
The first challenge we encountered in this project is the lack of readily available free data sources. At the initial stage, we found some great APIs provided by renowned institutions and companies (such as Psychsignal, Sentdex, Reuters) that can provide precise and well-formed sentiment scores regarding cryptocurrency-related tweets, news, reddit comments and analysis reports. However, all of these APIs are either restricted to be run on specific kernels or charged at a very high price. We had to give up such data and use other more accessible alternatives.

Another big challenge we have encountered is that we need to find an appropriate way to integrate price data, news data and social data together. For example, for each hour-level price, there might be dozens of news happened in the past several hours. We first need to find an optimal time interval to integrate such news, and also find an appropriate method to transform such news into numerical features.

Also, cryptocurrency prices are hard to predict in nature, not only because cryptocurrency prices are driven by many factors and affected by many other assets, but also because investors can be very irrational sometimes. In order to capture the effective price drivers, it requires domain knowledge to properly generate valid features for the model to learn. In this report, we will explore the possibility of using solely price and sentiment data to predict cryptocurrency prices and test if they are valid price drivers.

As we aim to build a real-time streaming system, we need to automate all the processes, which include data gathering, data preprocessing, model sampling, as well as sending the prediction data to our web front-end to draw dynamic plot. Building the whole streaming pipeline involves many techniques across many different programming language, which is quite challenging for us.

3 Data Science Pipeline



3.1 Data Preparation and Feature Generation

We collect our data mainly from CryptoCompare. Its API provides wide range of market data includes cryptocurrency trade data, order book data, blockchain data, social data and historical data. In this project, we used historical data, news data, and social data for feature extraction.

3.1.1 Historical Data

For historical data, its API returns daily, hourly and minute historical data, daily data at any given timestamp, daily average price based on hourly vwap and total daily and hourly exchange volume. Its free API only provides minute-level data for the past 7 days. Therefore, we use 20,000 hourly-level data(from 2017.01.01 to 2019.04.01) to train our model instead. The historical trading data provides Open/High/Low/Close prices and volumes information regarding different currency pairs.

Instead of feeding such prices into the model directly, we transform such prices into log returns. The reason is that, if we feed the price directly into the LSTM model, the model is likely to just use yesterday's price as the prediction result for next day's price, and provides meaningless results. In order to solve this drawbacks and retain as much information as possible, we transform the Close price into log returns (divide current Close price by Close price at previous timestamp, and then take log), and divide High and Low price by Open price to capture candlestick behaviors.

3.1.2 News Data

CryptoCompare's news API provides news feeds and articles from all major crypto news providers. For each article content (both title and body), we get a sentiment score by using NLTK Vader's pretrained sentiment intensity analyzer. For each hourly price, we look at all related news published in the previous 24 hours until that timestamp, and calculate a mean score for such news, and use the mean score as a feature. The sentiment score provides 4 intensities: positive, neutral, negative, and compound. In this project, we keep all four features.

3.1.3 Social Data

CryptoCompare's social API is able to provide hourly social status for the coin requested. It provides numerical features such as analysis page views, market page views, github repo stats, and facebook/twitter/reddit stats.

Instead of using such numerical features directly, we use the difference of such stats between each hour (i.e. increased/decreased amount of views, facebook likes, reddit comments within each hour).

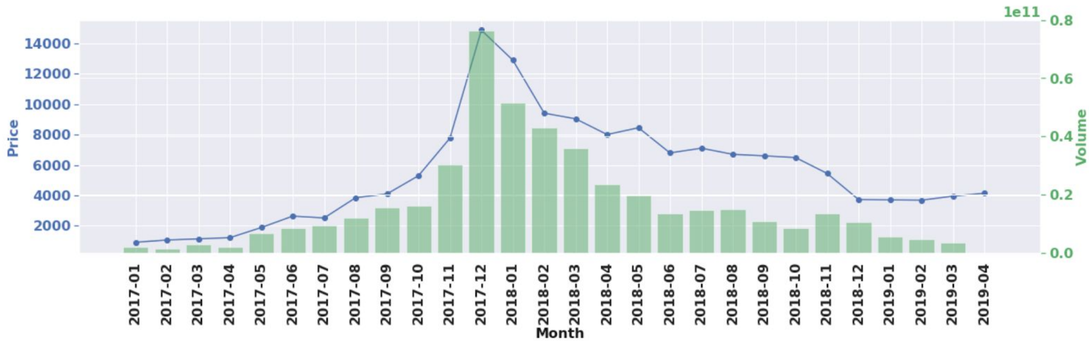
3.1.4 Missing Values

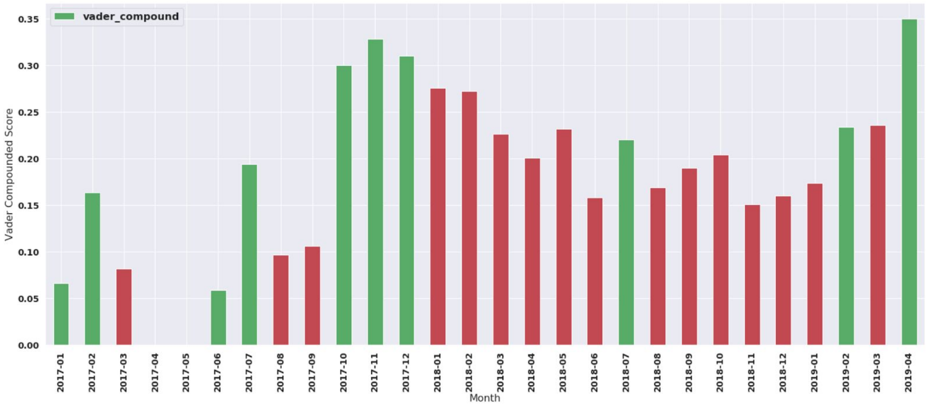
After aggregating and extracting above features, each piece of data has 42 features. For data in earlier time, some social data are missing. Since such social data are non-existent for early stages, it makes sense to simply fill such social missing values with zeros. Although special cases may still be present, in order to streamline the preprocessing and make it part of the pipeline, those were not currently dealt with.

3.2 EDA

After getting the market and social data from the API, we generated sentiment score according to the date the social data is released. Exploratory data analysis was carried out for the purpose of visualizing the trends of price and sentiment changes in the 27 months worth of data. Due to the volatility of the cryptocurrency market, although the covariance nature of price/volume and sentiment score is observed; however, the causation relationship is often reversed. Generally speaking, sometimes the public expectations drives the market while sometimes the market affect the general sentiment. Furthermore, very often a short interval is observed between the change of price/volume and sentiment since the change of one variable usually require some time for the other variable to realize.

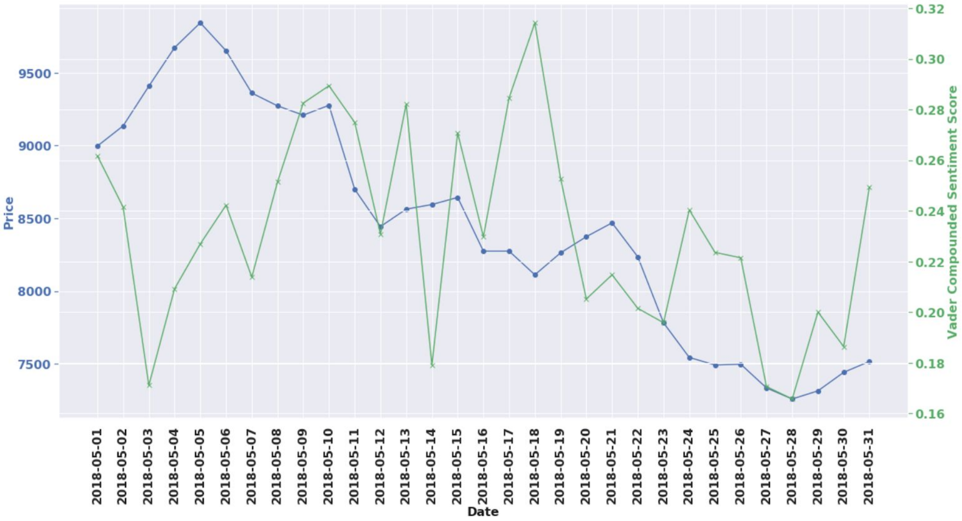
For example, the general price and market cap shown below clearly indicates the sudden spike of the popularity of the cryptocurrency market at the end of 2017.





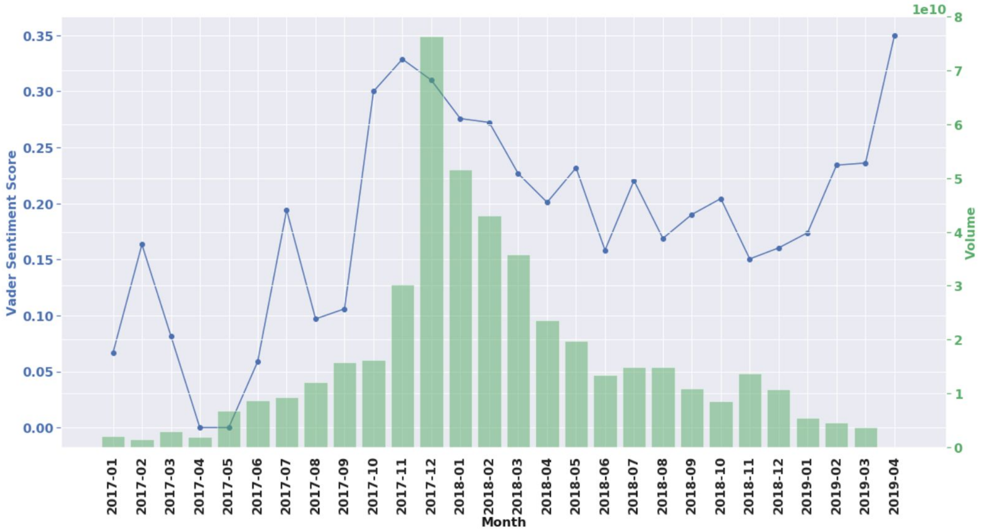
However, one could not help wondering that what the public sentiment is regarding such huge jump of price. Now we show the sentiment score obserserved. As shown in the above graph, green is considered positive expectation and red is the opposite. The corresponding sentiment scores of each green bar is either very high or increased significantly from previous month. One interesting observation is that as early as 2 months before the price spike, the public expectation about cryptocurrency is already starting to look forward towards a bull market and the accumulation of positive sentiment is perhaps one of the major factors that eventually leads to the price to reach its peak 2 months later.

However, as mentioned before, the intertwined relationship between price and sentiment score tells us that unlike the example where expectation drives the price, sometimes that public sentiment may also be affected by the price in similar fashion such as the graph shown below.



If one looks closely, on 3rd of May, 2018 the sentiment score dropped to a very low level but the price, on the contrary, stood strong and even increased for 2 consecutive days despite the general negative mood. Later, starts on May 4th, the sentiment score immediately bounced back from the previous trough which means that after seeing the strong position the price is holding, the public realizes that the future is perhaps more optimistic than they previously imagined.

Furthermore, as shown in the following graph, the market cap generally follows the sentiment score with several exceptions. This shows that the cryptocurrency market and the sentiment is definitely not in a monotonic relationship since that the amount of trading is not entirely depend on the sentiment of the public which shows the rationality of the traders. This will of course make the cryptocurrency market, like any other financial market, more complicate for people to understand.



3.3 Model Development

In this project, we use Keras with TensorFlow backend to implement LSTM. Long short-term memory (LSTM) is a type of recurrent neural network (RNN) that is able to recognize patterns in sequences of data and learn trends in time. Thus, we determine it to be suitable for financial time series data. In this report, we investigate the possibility of predicting future coin prices based on historical price, sentiment score, and social stats.

3.3.1 Time Series to Supervised Learning

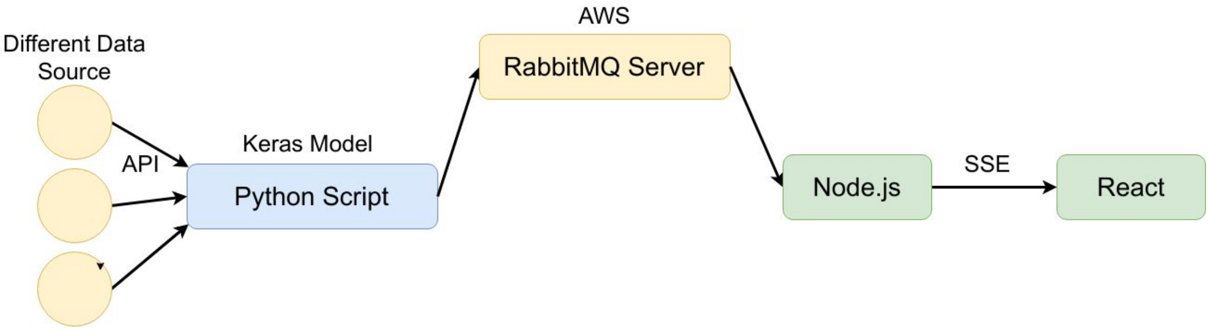
For Keras, we need to preprocess the Time Series data and transform it into supervised learning. We first use sklearn MaxAbsScaler to normalize the data between 1 and -1. The reason why we use MaxAbsScaler instead of MinMaxScaler is that, by using MaxAbsScaler, our normalized data can still keep the positive and negative sign of return. After normalization, for each hourly timestamp, we use a sequence length of 50, and transform each 50 pieces of Time Series data into 1 piece of training/testing data. This means we use return/sentiment/social information at time T-49, T-48…, T-1 to predict the return at time T. After these steps, each feature vector is a numpy array with dimension (49, 42) (49 is the time window, 42 is the number of actual features).

3.3.2 Model Structure

Our model has 3 bi-directional hidden layers with 100 neurons each, and dropout ratio of 0.3. Since we are doing regression instead of binary classification, we use linear as activation function. The training is performed on the split (0.8) train data in batches of size 1024, using Adam optimizer and mean-squared-error as the loss function. We trained the model for 100 epochs.

Our model will output predicted log return. After denormalization and transformation to actual coin price, we feed the result to our real-time system.

3.4 Real-Time System Pipeline-Hoppe



Our streaming pipeline structure is shown above. Since we can only calculate hour-level sentiment data, we built a streaming system embedding a LSTM RNN model trained with past hour-level data to predict the price of Bitcoin for the next hour.

Every hour, our Python script gathers the hour-level data: historical price data, news data, and twitter data, though CryptoCompare API, transform it and calculate the sentiment score based on news and twitter data. With further processing, such as data aggregation and data normalization, the data is then feed to the Keras model we trained to make prediction. Since we are predicting the return instead of the actual price, we need to transform the return back into price data.

After getting the predicted price for the next hour, we sent the data to one of the queues of the RabbitMQ server deployed on AWS. Our Node.js served as back-end server for our web application subscribed the queue we are sending message, once it received a new data point, it sent a SSE (Server Send Event) to our React front-end. On the front-end, we can then dynamically plot the price line plot based on the received event.

Because the python script sent message on hourly base, it's hard to see the dynamic update of our plot on the web application, we wrote another message sender which use the same model architecture but trained with random data to generate second-level data. We sent the fake data to another queue of our RabbitMQ server, so our plot on the front-end is able to draw new points every second, meaning that if we can get second-level sentiment data or historical price data, our system is able to achieve second-level streaming.

3.5 Web Application

Our website is not only a real-time cryptocurrency price prediction platform, but can also provide latest market information and statistical analysis to the users.

React is used to build the frontend of our platform. Since our platform is designed to be a dynamically updated(seconds-level) website, React would be a perfect choice as it would only update the components needed to be updated rather than rendering the whole page every time.

Node.JS is used to serve as our backend server in order to receive the real-time data from RabbitMQ located on AWS. Basically, our nodejs server would receive real-time data and send the data to the frontend when required.

Bootstrap is included to support our components' styling. This popular and lightweight framework can also provide customizable components to make our website more responsive and flexible.

HighChart is Javascript library that is used to plot our real-time price prediction graph updated in every second.

4 Methodology

4.1 Cryptocurrency Price Prediction with Deep Learning

Recurrent Neural Network is very suitable for processing time series data with its recurrent structure: it pass a hidden state from current time step of computing to the next time step, representing the the information it remembered which might be useful for next step. Since we are predicting the Bitcoin price of next time step using historical Bitcoin price data, it's natural to use RNN to learn from this time series data and make prediction. In this way, our RNN model will be a sequence-to-label model: input a sequence of past time Bitcoin data and output a number which represents the the price of Bitcoin for next time step.

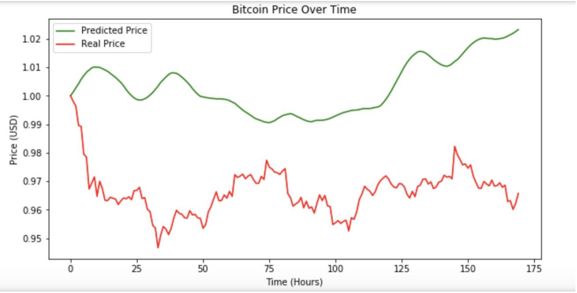
4.2 Real-time Pipeline with RabbitMQ

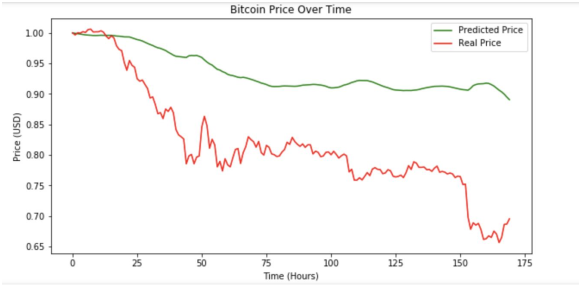
We used RabbitMQ as our streaming message broker because it's lightweight and easy to deploy. We deployed it on an AWS instance, so that it's able to receive and forward message from anywhere in the Internet. It received the real-time data from the python script running on our lab's computer and sent the data to the NodeJS server which was running on our own laptop.

5 Evaluation-Michael

We use 2 ways to evaluate our model performance.

First we compare our predicted return with actual return for our testing data. However, since we are predicting returns, one slight error (even 1%) will give dramatic difference with the addition of time. Therefore, instead of looking at the performance over the entire testing data (4000 hours), we divide our testing data into several chunks, with each chunk representing 168 hourly data points (i.e. 7 days). For each chunk, we reset the benchmark as 1, and evaluate the difference between predicted return and actual return for the next 168 hours.



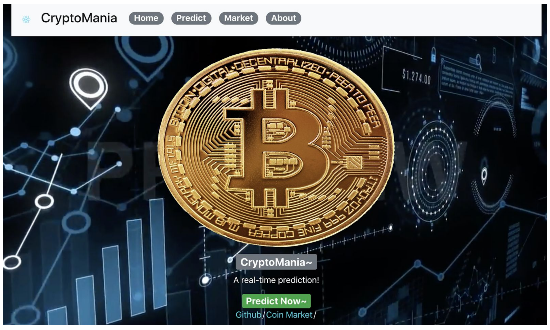


As shown by the graph, our model is able to capture some of the general trend of the coin return. The result is promising, and we believe that with more digging into hyperparameter tuning and more training time, we will be able to further improve the model performance.

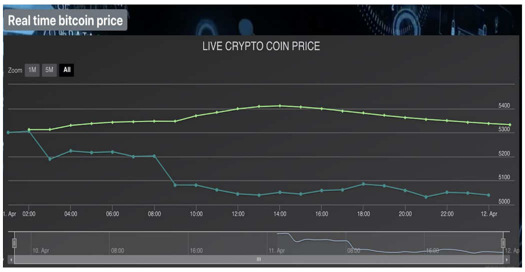
In addition to that, we also convert our predicted return into binary scores for evaluation. Specifically, if a predicted return is larger than 1, this means an predicted increase, then we mark it as 1. On the contrary, we mark a predicted decrease as 0. We do the same conversion for label return, and calculate the accuracy. Our model is able to provide an accuracy of 52%. Although this may sound low, we will even be able to make a profit on the long run with accuracy larger than 50%, given coin market is highly stochastic. We also believe that with further tuning, our model can achieve better performance.

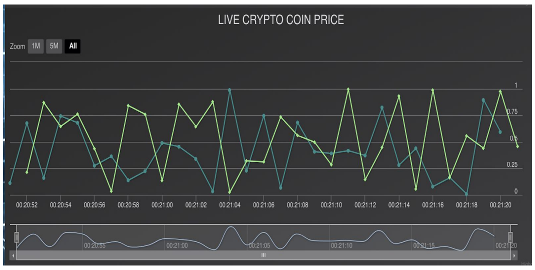
6 Data Product-Frank

Our data product is a web application that can receive real-time data from different sources and update the platform dynamically.

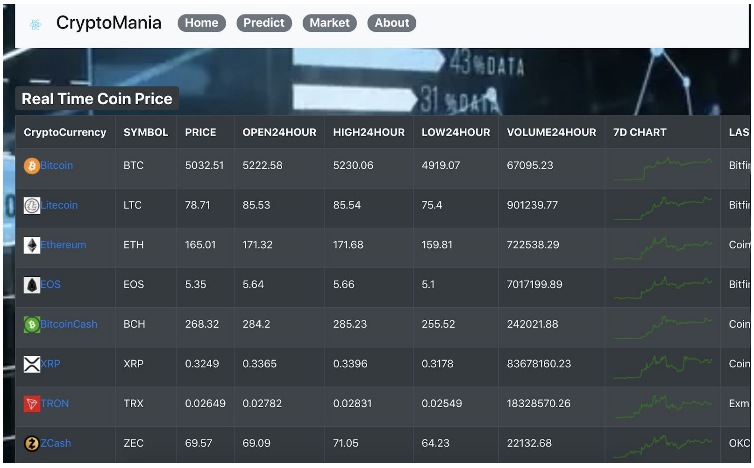


The above image is our home page. Our background is a cool video that is present through all web pages. The top white part is the navigation bar which can lead users to different pages.





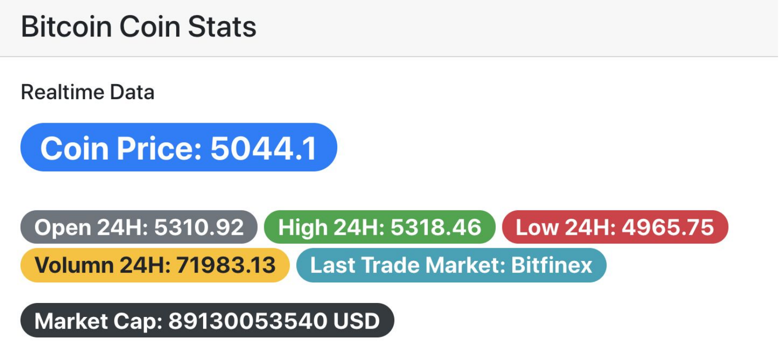
When you click 'Predict Now' button on the Home page, you will be directed to the prediction page. On the prediction page, on the first plot, the blue line is the actual Bitcoin price over the last 24 hours, and the green line is the price predicted by our model. We can see the predicted price is always one hour ahead of the actual price. This plot update once an hour, since we can only get hour level data and our model was trained on that. In order to test our streaming system, we plot the second plot with random data as input for our model, instead of making prediction once an hour, we make prediction every second. As we can see, the second plot dynamically updated every second, indicating that if we have second-level Bitcoin historical data, our system can achieve second level prediction streaming.



Real Time Coin Price

CryptoCurrency	SYMBOL	PRICE	OPEN24HOUR	HIGH24HOUR	LOW24HOUR	VOLUME24HOUR	7D CHART	LAS
Bitcoin	BTC	5032.51	5222.58	5230.06	4919.07	67095.23		Bitfi
Litecoin	LTC	78.71	85.53	85.54	75.4	901239.77		Bitfi
Ethereum	ETH	165.01	171.32	171.68	159.81	722538.29		Coin
EOS	EOS	5.35	5.64	5.66	5.1	7017199.89		Bitfi
BitcoinCash	BCH	268.32	284.2	285.23	255.52	242021.88		Coin
XRP	XRP	0.3249	0.3365	0.3396	0.3178	83678160.23		Coin
TRON	TRX	0.02649	0.02782	0.02831	0.02549	18328570.26		Exm
ZCash	ZEC	69.57	69.09	71.05	64.23	22132.68		OKC

At market page, we select 8 most popular cryptocurrencies. For each coin, we present 10 different features of the coin. The whole table is automatically updated since we use an API to fetch required data every second. So the table would present the latest statistics of these cryptocurrency.



Bitcoin Coin Stats

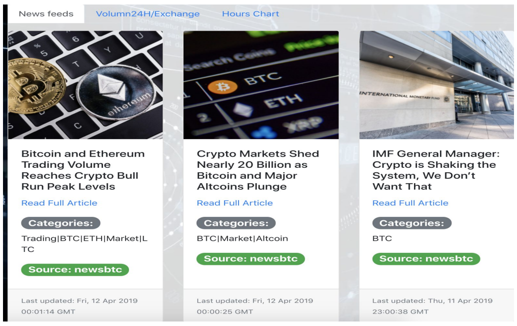
Realtime Data

Coin Price: 5044.1

Open 24H: 5310.92 High 24H: 5318.46 Low 24H: 4965.75
Volumn 24H: 71983.13 Last Trade Market: Bitfinex

Market Cap: 89130053540 USD

When clicked on a certain cryptocurrency, the user will jump to the coin details page. All the coin statistics data is real-time and they are updated every seconds.



Bitcoin and Ethereum Trading Volume Reaches Crypto Bull Run Peak Levels
Read Full Article
Categories:
Trading|BTC|ETH|Market|LTC
Source: newsbtc
Last updated: Fri, 12 Apr 2019 00:01:14 GMT

Crypto Markets Shed Nearly 20 Billion as Bitcoin and Major Altcoins Plunge
Read Full Article
Categories:
BTC|Market|Altcoin
Source: newsbtc
Last updated: Fri, 12 Apr 2019 00:00:25 GMT

IMF General Manager: Crypto is Shaking the System, We Don't Want That
Read Full Article
Categories:
BTC
Source: newsbtc
Last updated: Thu, 11 Apr 2019 23:00:38 GMT

On the details page, we also fetch the three latest news that are related to this specific cryptocurrency. These news are updated every 10 minutes automatically. Users can also click the links to see the full article, and see the categories and source of the news.

7 Lessons Learned

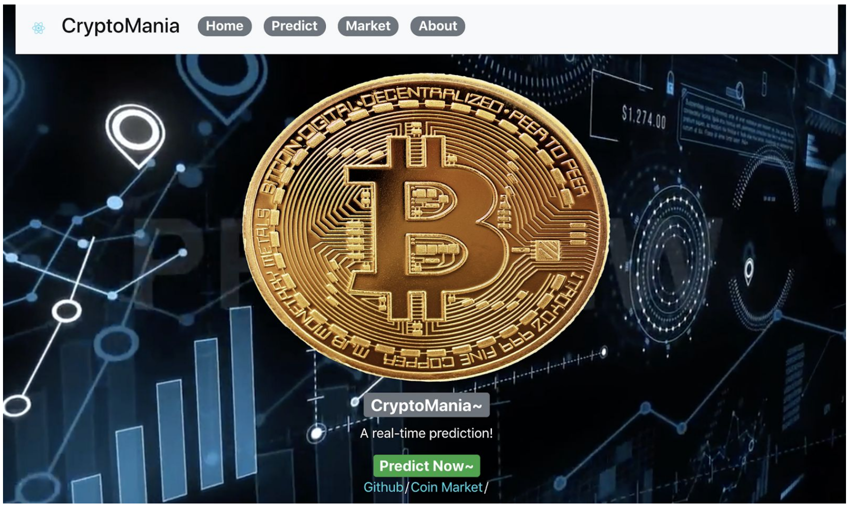
In this project, we learned how to solve ambiguous problems and build a data product from scratch. Different from other course projects or assignments where we were given very specific tasks, in this project, we were only given a very broad topic to work on. We need to find our own datasets, explore data information and do feature generation with creativity, build our own pipeline, find the most suitable model, identify specific target groups, and build a promising data product that can make an impact.

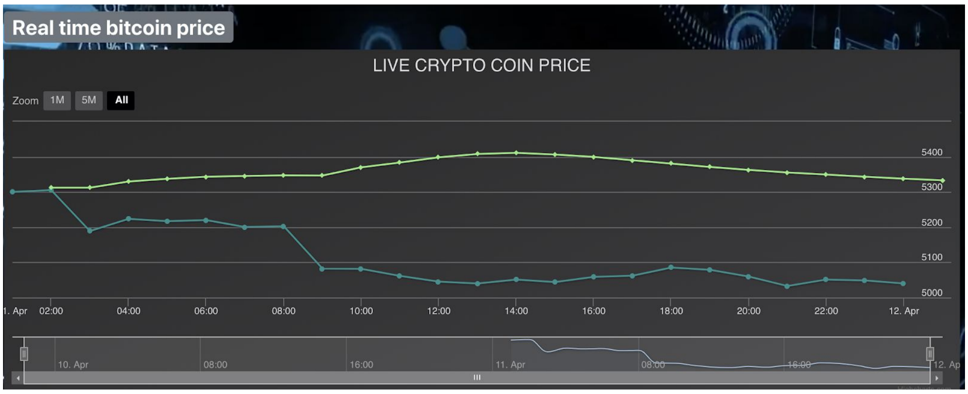
Technology-wise, we learned how to integrate data from different resources, how to transform raw time series data into supervised learning, and how to build a robust real-time system with new technologies. We also learned how to build a comprehensive and robust web application with appealing and customizable frontend using React, and connect it with our long-running backend server written by Nodejs.

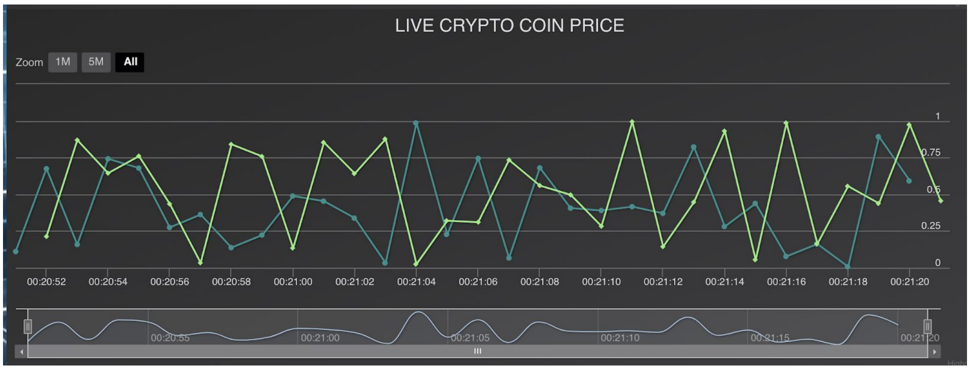
8 Summary

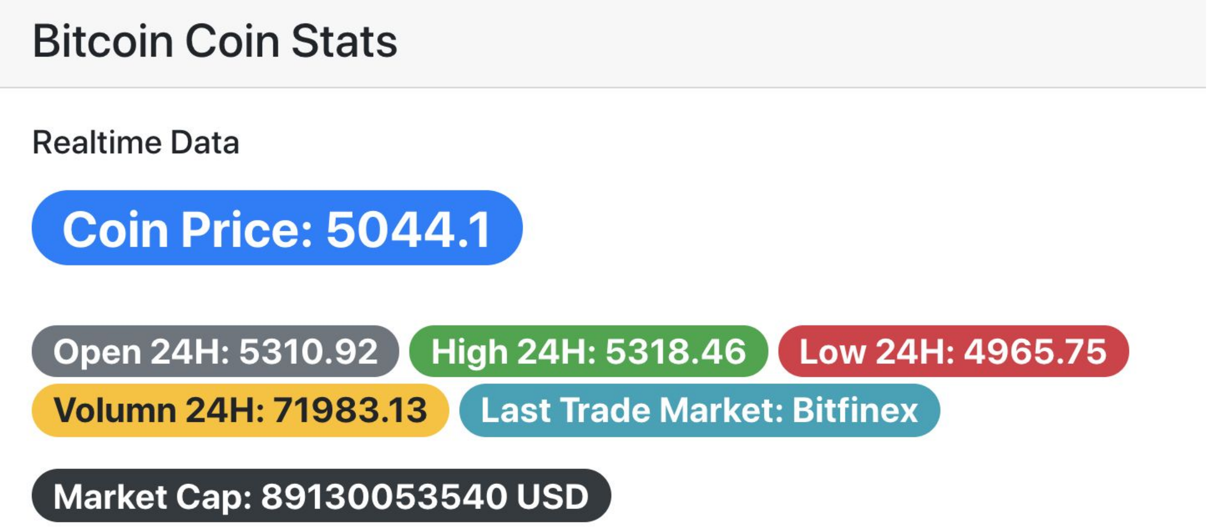
In this project, we build a one-stop web application for cryptocurrency enthusiasts. By fetching our data through API, our web application is able to provide ever-updating, comprehensive information regarding cryptocurrencies in all aspects. By integrating cryptocurrency price data with news sentiment analysis and social status information, we build a deep learning model that is able to provide predictions on coin prices or binary returns to help investors with their decisions. By further integrating the model with our real-time pipeline, our web application is able to provide dynamic prediction curve every hour, minute or even second.

Appendix


CryptoMania
Home
Predict
Market
About
$1,274.00
CryptoMania~
A real-time prediction!
Predict Now~
Github/Coin Market/


Real time bitcoin price
LIVE CRYPTO COIN PRICE
Zoom
1M
5M
All
5400
5300
5200
5100
5000
1. Apr
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
12. Apr
10. Apr
08:00
16:00
11. Apr
08:00
16:00
12. Ap


LIVE CRYPTO COIN PRICE
Zoom
1M
5M
All
1
0.75
0.5
0.25
0
00:20:52
00:20:54
00:20:56
00:20:58
00:21:00
00:21:02
00:21:04
00:21:06
00:21:08
00:21:10
00:21:12
00:21:14
00:21:16
00:21:18
00:21:20
00:20:55
00:21:00
00:21:05
00:21:10
00:21:15
00:21:20



Bitcoin Coin Stats

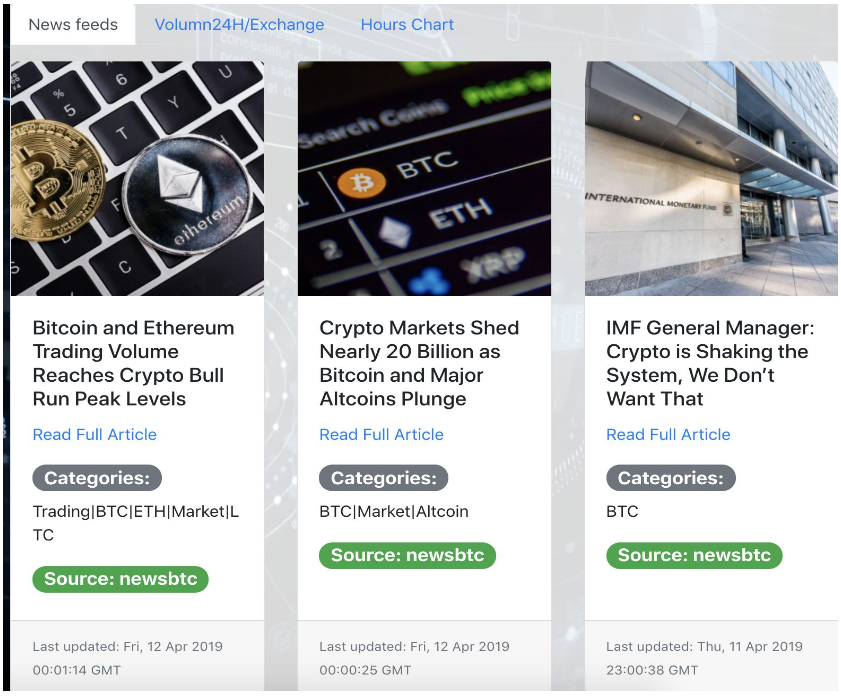
Realtime Data

Coin Price: 5044.1

Open 24H: 5310.92 **High 24H: 5318.46** **Low 24H: 4965.75**

Volumn 24H: 71983.13 **Last Trade Market: Bitfinex**

Market Cap: 89130053540 USD



Bitcoin and Ethereum Trading Volume Reaches Crypto Bull Run Peak Levels

Read Full Article

Categories:

Trading|BTC|ETH|Market|LTC

Source: newsbtc

Last updated: Fri, 12 Apr 2019 00:01:14 GMT

Crypto Markets Shed Nearly 20 Billion as Bitcoin and Major Altcoins Plunge

Read Full Article

Categories:

BTC|Market|Altcoin

Source: newsbtc

Last updated: Fri, 12 Apr 2019 00:00:25 GMT

IMF General Manager: Crypto is Shaking the System, We Don’t Want That

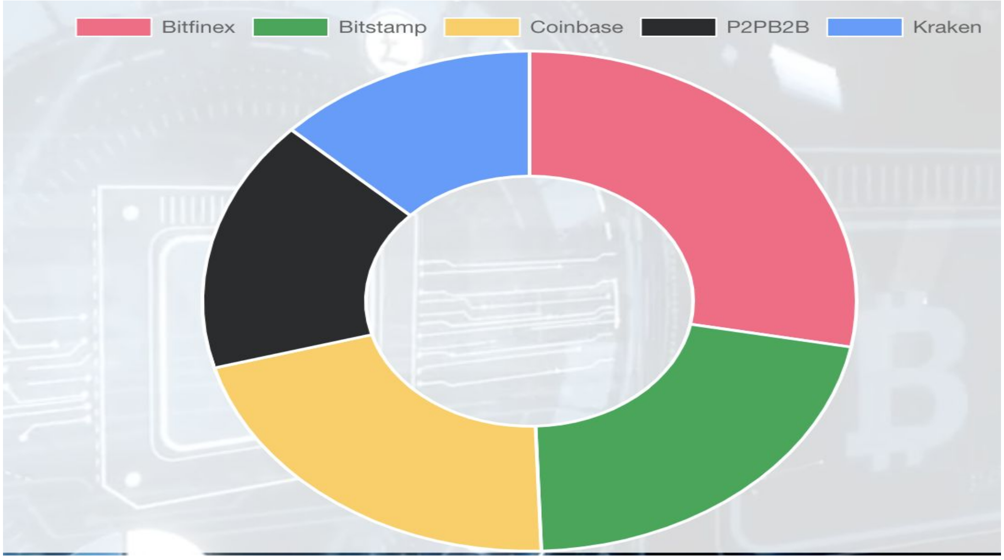
Read Full Article

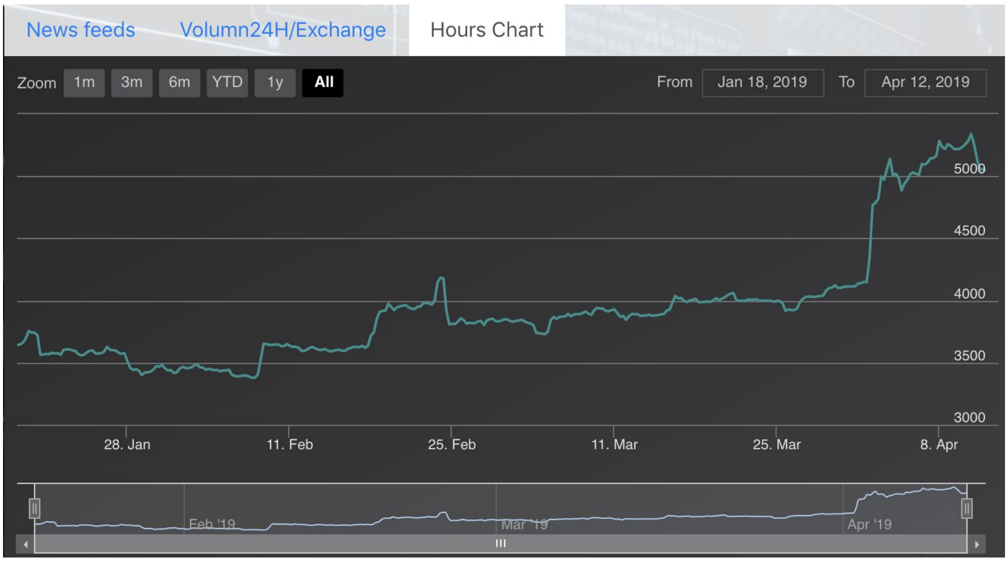
Categories:

BTC

Source: newsbtc

Last updated: Thu, 11 Apr 2019 23:00:38 GMT





From Jan 18, 2019 To Apr 12, 2019