Machine Learning Engineer Nanodegree

Capstone Proposal

Tony Ng
August 17, 2019

Domain Background

Bitcoin is a digital currency established in 2009 by Satoshi Nakamoto. It is decentralized and transactions are verified by network nodes through cryptography and recorded in a public distributed ledger called blockchain [1]. More and more products and services can be paid by bitcoin in recent years, and exchange trading volumes continue to increase. Poloniex, a digital asset trading service, saw an increase of more than 600% active bitcoin traders online and regularly processed 640% more transactions between January and May 2017 [2].

Bitcoin is adopted worldwide; however, its price volatility has also widely been criticized. One bitcoin was worth less than 3000 US dollars at the start of 2017, but it jumped to around 20000 US dollar at the end of that year. A year later, the price fell to around 3000 US dollars again. According to Mark T. Williams, bitcoin has volatility seven times greater than gold, eight times greater than the S&P 500, and 18 times greater than the US dollar [3]. People cannot predict the bitcoin price now, so a reliable, bitcoin-price prediction tool should be implemented to reinforce the confidence of people to bitcoin.

Problem Statement

The problem is to predict the bitcoin price on the next day based on its historical price data. A machine learning model can be used to learn the pattern of bitcoin historical price data and makes inference to the future price.

Datasets and Inputs

The dataset includes one CSV file that contains minute-to-minute bitcoin exchange data recorded by Bitstamp from January 1, 2012 to August 12, 2019 [4]. The file contains 3997697 records and have eight columns:

- Timestamp - Start time of time window (1 minute) in UNIX time
- Open - Open price at start time window
- High - High price within time window
- Low - Low price within time window
- Close - Close price at the end of time window
- Volume_(BTC) - Amount of BTC transacted in time window
- Volume_(Currency) - Amount of Currency transacted in time window
- Weighted_Price - Volume-weighted average price

Time windows without any trade have their data fields (except timestamp) filled with NaNs. The dataset records will be grouped by daily basis and the volume-weighted average price of bitcoin on each day between January 1, 2012 to August 12, 2019 will be calculated. This volume-weighted average price will represent the price of bitcoin on a certain day and acts as the historical price data of bitcoin to train up the machine learning model to make predictions.

Solution Statement

We will use time series forecasting to predict the bitcoin price on the next day. We divide the dataset into training and testing sets: the first 80% of the data in training and the last 20% in testing. First, the model is trained on the training set and predicts the bitcoin price on the first day of the testing set. Then, we add the price on the first day of the testing set to the training set, train the model and predict the price of the second day. We repeat this procedure on all days in the testing set. As a result, the true values and the predicted values on the testing set can be compared to evaluate the performance of the model.

Benchmark Model

We will use univariate analysis on ARIMA model as the benchmark model. ARIMA is easy to implement and linear, so it is good to act as a baseline to compare to other complicated models. We can use the procedure stated in the ‘Solution Statement’ section to evaluate the performance of ARIMA model.

Evaluation Metrics

We will use R2 score to quantify the performance of benchmark model and the solution model. In regression, the R2 score is a statistical measure of how well the regression predictions approximate the real data points. An R2 of 1 indicates that the regression predictions perfectly fit the data. And the R2 score formula is calculated by dividing the sum of the errors of our model by the sum of the errors of the simplest possible model and subtracting the derivation from 1.

Project Design

Firstly, we will do data cleaning and drop out any record in the dataset that contains fields with the NaN value. Then we will group the data on a daily basis and calculate the volume-weighted average price on each day. The data will be split into training and testing set. An ARIMA benchmark model will be trained and evaluated on the performance by the testing set.

Next, we will perform feature engineering to create new features including moving average of past 5/10/30 days and historic volatility. Correlation analysis on the new features will find out which features are closely related. Not closely related features will be used and new feature values will be calculated on each data points.

After the dataset is ready, we train up a Vanilla LSTM and a stacked LSTM. Finally, we compare and discuss the performance of the LSTM models and the benchmark model by calculating their R2 scores. The one having the best R2 score value will be chosen as the solution model of this project.

References

[1] Bitcoin -Wikipedia
https://en.wikipedia.org/wiki/Bitcoin

[2] Bitcoin History – Price since 2009 to 2019, BTC Charts – BitcoinWiki
https://en.bitcoinwiki.org/wiki/Bitcoin_history

[3] Virtual Currencies - Bitcoin Risk
Mark T. Williams
http://www.bu.edu/questrom/files/2014/10/Wlliams-World-Bank-10-21-2014.pdf

[4] Bitcoin Historical Data | Kaggle
https://www.kaggle.com/mczielinski/bitcoin-historical-data