ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and
Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1.
Available from http://www.i6doc.com/en/.

Price direction prediction in financial markets, using Random Forest and Adaboost

Mohammadmahdi Ghahramani[1] and Fabio Aiolli[1]

1- Università di Padova, Department of Mathematics - Padova, Italy

Abstract. Experience shows trading in financial markets can be highly profitable. In this light, a great deal of effort has been devoted to using machine learning to predict market behavior. By using Random Forest and Adaboost models, we present a novel method for modeling candlestick patterns in financial markets. Our first contribution in the preprocessing part is to prepare data, develop additional features, and modify data. Our second contribution is introducing a novel prediction approach, named dataset ensembling to predict daily prices. Using three-year daily Bitcoin prices, the models are trained, tuned and then tested on one year of unseen data, showing the feasibility of the approach in terms of accuracy.

Introduction

Trading cryptocurrency has recently gained popularity. Researchers have been seeking ways to predict financial markets. Methods range from traditional machine learning models[1, 2] and simple neural networks [3] to complex deep learning based models. Complex models mostly take advantage of Recurrent Neural Networks (RNNs) [4] and due to long dependencies in financial markets, Long Short Term Memory (LSTM) units are widely used [5, 6]. Price direction prediction using classification [7] and regression [8] are types of approach in recent studies. This study proposes a simpler and less expensive model, compared to state-of-the-art models such as LSTM, to determine price movement direction.

A candlestick pattern is a common pattern in technical analysis that helps traders anticipate price direction. The first part of this study involves preparing data in order to model these patterns. Afterward, it discusses the concept of being *data stationary* and makes an effort to make the data stationary.

The machine learning techniques sometimes use many models to simultaneously predict outputs, thereby reducing variance. The second part of this paper, modeling part, instead of using multiple models, suggests using multiple datasets to estimate data samples' target value. It tests nine different models and discusses their performances. The models might try to directly classify the price direction or to first perform a regression task and extract price directions using outputs. Dataset used in this research is gathered according to Bitcoin daily price from 2017 to 2021.

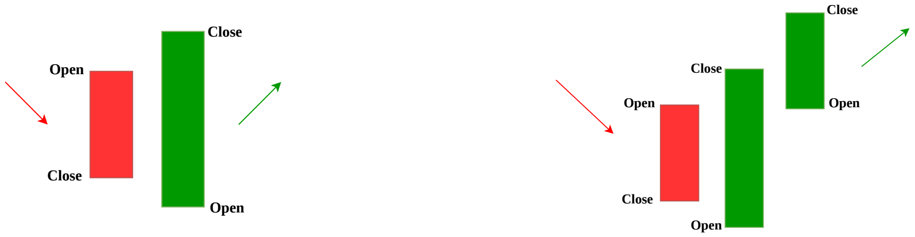
1 Data design

In financial markets, it is common to picture the price fluctuation using candles. There are generally five characteristics considered for each candle. They are

Open, High, Low, Close prices and transactions' Volume (OHLCV). Candles might be built over a range of time frames, from five minutes to yearly. This part introduces candlestick patterns and suggests a set of features by which these patterns can be modeled. Then the concept of data stationary is discussed.

1.1 Candlestick patterns modeling

From an economic point of view, there is a strong link between price and supply-demand pressure. However, it has been discovered that the markets are strongly influenced by the emotions of traders. Candlesticks show that emotion by visually representing the size of price moves with different colors. Traders use the candlesticks to make trading decisions based on regularly occurring patterns that help forecast the short-term direction of the price. Most of these patterns include one, two or three consecutive candles. However, those patterns with three candles are more accurate and robust. Figure 1 shows two different candlestick patterns, consisting of two and three candles.



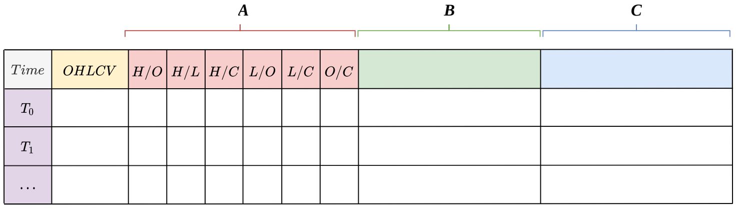
(a) Bullish engulfing

(b) Three outside up

Fig. 1: **Examples of candlestick patterns.** a) It consists of two candlesticks, with the second candlestick engulfing the first candlestick. The first candle is bearish, indicating the downtrend will continue. It shows that the price goes up. b) It consists of three candlesticks, the first being a short bearish candle, the second candlestick being a large bullish candle which should cover the first candlestick. The third candlestick should be a long bullish candlestick confirming the bullish reversal.

To model such patterns, this paper suggests generating additional features apart from OHLCV. Patterns are mainly about price pressures and can be summed up as follows: if a specific ratio between OHLC prices is provided then the price moves in a particular direction. Taking advantage of this observation, the study considers six more features for each candle, including the pairwise ratio between OHLC prices.

Most stable patterns are made of three candles. For each candle, we consider 18 features, six features for the current candle, six features for the previous candles, and six features for the two previous candles. Apart from pattern modeling, these features are capable of bringing the short-term memory from previous candles, without using RNNs. Figure 2 provides an illustration of this mechanism.



Time	*OHLCV*	*H/O*	*H/L*	*H/C*	*L/O*	*L/C*	*O/C*		
T_0									
T_1									
…									

Fig. 2: **New generated features.** A, B and C corresponds to the current, previous and two previous candles.

1.2 Stationary analysis

In time series analysis, data is needed to be stationary. It means that the mean and variance of data should remain unchanged as time passes. This makes the training and test data similar in distribution. Hence, it is crucial to check if the time series data we use is stationary. In most cases, this assumption is not held, because there are price trends in financial markets.

One common way to make the data stationary, is to use first order differentiation between prices. However, using this trick, we lose data interpretability, meaning that the generated data cannot show the fluctuations in the original data. This study refers to this issue as memory vanishing.

To reduce the effect of memory vanishing and still having stationary data, the *fractionally differentiation* is suggested by this study [9]. In first order differentiation we simply use two consecutive data. However, in the fractionally version, we first compute a set of coefficients and then apply them on several previous prices to produce the output. The number of coefficients is a hyperparameter to set. Coefficients are defined as the following equation, where `w` is a set of coefficients and a real positive `d` preserves memory. This study uses `d`=$\frac{1}{2}$.

$$w = \{1, -d, \frac{d(d-1)}{2!}, -\frac{d(d-1)(d-2)}{3!}, ..., \frac{(-1)^k}{k!}\prod_{i=0}^{k-1}(d-i), ...\} \tag{1}$$

The outputs then are computed as below, where $\tilde{X_t}$ is a more stationary version of X_t that has memory. θ determines how many previous data points contribute to producing the output of the current data point. This study uses θ=10. By setting θ=0, we do not apply any modification on data and `d`=θ=1 is equivalent to the first order differentiation.

$$\tilde{X_t} = \sum_{k=0}^{\theta} w_k X_{t-k} \tag{2}$$

2 Modeling

In this part, firstly the concept of *dataset ensembling* is introduced. Then nine different models are trained, fine tuned and tested on data. Models may differ in features, methods, or approaches that they use to obtain the output.

2.1 Dataset ensembling

One of the most informative fluctuations in the market is the daily fluctuation of the price. The daily candle is closed at 00:00 every night. There are several other time frames that coincide with this time. As an instance, candles with the time frame of eight-hour coincide with 00:00 every three times.

This paper uses five different datasets, including those with the time frame of 30-minute, 1-hour, 2-hour, 4-hour and 8-hour. Figure 3 shows how they produce five different estimations of daily close prices. If the task is classification, the paper uses majority voting and if the task is regression, it uses a simple averaging between five produced outputs to decide the final output.

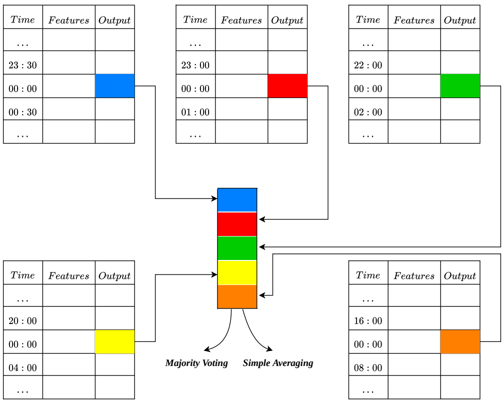


Fig. 3: **Dataset ensembling**

2.2 Model selection and evolving

This section introduces nine models for price direction prediction tasks. By testing these models, the paper checks three facts. The first one is if candlesticks patterns of three candles are really more accurate than those with two or one candles. The second one is if changing the approach from classification to regression, with remaining the method unchanged, helps us improve the model accuracy. Finally it checks whether changing the method makes the output behave better. Two methods, Random Forest and Adaboost, as well as two approaches, direct classification and regression with trading strategy, form our nine models.

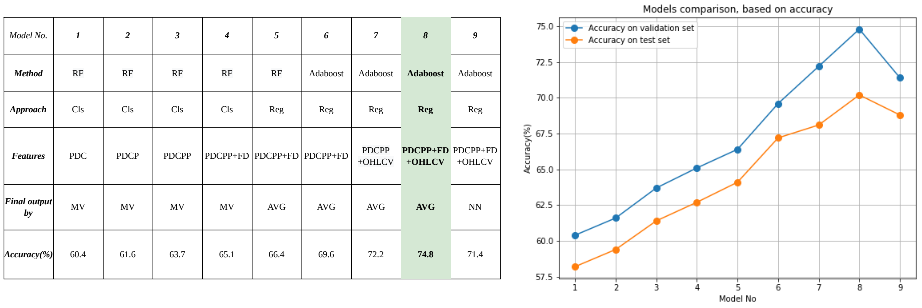
The intuition that candlestick patterns can be modeled by defining a rule-based system, requires the paper to use a tree-based model. That is why the first method is Random Forest. The second method, Adaboost, is quite robust to overfitting and it makes sure that the complexity of data and model does not result in poor accuracy on unseen data.

In the case that we use direct classification, we simply compute the accuracy between the model's output and ground truth labels, while in the case of regression with trading strategy, we first predict the close price and by using a trading strategy we measure how many of times the direction predicted by our predicted close prices is the same as what happens in reality. We convert continuous values for close prices to the discrete values of 0 and 1 according to the following equation and compare the result with the real fluctuation in the market, where $\hat{y}_i$ is the prediction model makes for the `i-th` candle, using regression.

$$\text{TradingStrategy(i)} = \begin{cases} 1 \text{ (up)} : & \text{if } \hat{y}_{i+1} - \hat{y}_i \geq 0 \\ 0 \text{ (down)} : & \textit{Otherwise} \end{cases} \tag{3}$$

3 Results

Fig 4 consists of a table reporting accuracy of each model on the validation set and a figure visualizing the models' accuracy on the validation and test set. The result indicates that using more previous candles, we get more accurate.



Model No.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***
Method	RF	RF	RF	RF	RF	Adaboost	Adaboost	**Adaboost**	Adaboost
Approach	Cls	Cls	Cls	Cls	Reg	Reg	Reg	**Reg**	Reg
Features	PDC	PDCP	PDCPP	PDCPP+FD	PDCPP+FD	PDCPP+FD	PDCPP +OHLCV	**PDCPP+FD +OHLCV**	PDCPP+FD +OHLCV
Final output by	MV	MV	MV	MV	AVG	AVG	AVG	**AVG**	NN
Accuracy(%)	60.4	61.6	63.7	65.1	66.4	69.6	72.2	**74.8**	71.4

(a)

(b)

Fig. 4: **Model testing.** Considered methods are Random Forest (RF) and Adaboost. The set of approaches contains direct classification (Cls) and regression with trading strategy (Reg). Regarding features, PDC stands for Pairwise Division for the Current candle, PDCP stands for Pairwise Division for the Current and Previous candles, PDCPP stands for Pairwise Division for the Current and two Previous candles and FD stands for Fractionally Differentiated version of OHLCV. Depending on the approach, Majority Voting (MV) or Simple Averaging (AVG) is used to produce the final output. NN also indicates, instead of averaging or voting, there is a neural network in charge of producing the final output.

Conclusion

This research provides a novel way to model candle stick patterns in financial markets, using dataset ensembling approach. It starts with defining new features and then attempting to make the time series data stationary. It tests nine models on provided data. The result endorses the fact that even without complex architectures, such as LSTM, we are able to bring a short memory in our model and have an acceptable error of prediction. From a trading perspective, this accuracy combined with expertise can result in substantial gains. Since profit in the financial market is more important than accuracy, further research can be considered to measure and optimize the amount of profit gained by the models.

References

[1] Shunrong Shen, Haomiao Jiang, and Tongda Zhang. Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, pages 1–5, 2012.

[2] Bruno Miranda Henrique, Vinicius Amorim Sobreiro, and Herbert Kimura. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124:226–251, 2019.

[3] Matthew Dixon, Diego Klabjan, and Jin Hoon Bang. Classification-based financial markets prediction using deep neural networks. *Algorithmic Finance*, 6(3-4):67–77, 2017.

[4] Nijolė Maknickienė, Aleksandras Vytautas Rutkauskas, and Algirdas Maknickas. Investigation of financial market prediction by recurrent neural network. *Innovative Technologies for Science, Business and Education*, 2(11):3–8, 2011.

[5] Adil Moghar and Mhamed Hamiche. Stock market prediction using lstm recurrent neural network. *Procedia Computer Science*, 170:1168–1173, 2020.

[6] Mohammadmahdi Ghahramani and Hamid Esmaeili Najafabadi. Compatible deep neural network framework with financial time series data, including data preprocessor, neural network model and trading strategy. *arXiv preprint arXiv:2205.08382*, 2022.

[7] Nagaraj Naik and Biju R Mohan. Stock price movements classification using machine and deep learning techniques-the case study of indian stock market. In *International Conference on Engineering Applications of Neural Networks*, pages 445–452. Springer, 2019.

[8] Wei Bao, Jun Yue, and Yulei Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944, 2017.

[9] Marcos Lopez De Prado. *Advances in financial machine learning.* John Wiley & Sons, 2018.