**Machine Learning for Algorithmic Trading**

**Abstract**

In this project, we made an effort to predict the trend of crypto using machine learning algorithms. For early stages of analysis, we intended to comprehend and better recognize daily crypto market movements, as well as acquire insight into ideal crypto pricing features. Our data set contains more than five factors related to the current crypto price. In addition, we extracted some features from the dataset using a variety of external libraries and different calculations. With this information and three distinct machine learning algorithms, we were 91% accurate in forecasting the pattern of daily market moves. As the second part of our analysis, we modified our dataset from candlestick chart form to renko chart pattern and were 96% efficient in anticipating the pattern of daily market moves.

**Introduction**

Literature Review

1. For many decades, researchers have been interested in determining the key causes of stock returns. Stock returns, as demonstrated in traditional financial theories such as CAPM and numerous multivariate models (Fama and French, 1993[4], 2015[4]), are linearly assigned to underlying factors such as systematic risk, market size, book to market ratio, and so on.
2. The accompanying linear regression scheme, in conjunction with the substantially established variables spanning many technical and basic features (Zhu et al., 2011[6]), serves as the major workhorse for financial modelling in the academic and industrial sectors. Yet, it is unclear if the market is linear and whether the return is linearly repressible or solely due to market anomalies (Zhu et al., 2012[7]).
3. To forecast the stock movement, several researchers have employed support vector machine, random forest, and other machine learning models. Integration models have played an essential role in a variety of application domains [8-10]. Yet, dealing with time series data, choosing technical indicators, and optimizing the combination of factors remains difficult. Few research studies create stock forecasting models consistently based on this entire process; hence, it is critical to apply the technique described in this study to stock market forecasting.

Cryptocurrency

Crypto is a decentralized digital currency and payment system based on peer-to-peer transactions with no official control. Instead, encryption is used to secure network transactions and liquidity. The system was initially introduced in 2009 and has evolved into a thriving open-source community and payment network. The crypto ecosystem is getting a lot of interest from businesses, users, and entrepreneurs as a result of crypto's financial tool and its rapid popularity. To be explicit, in order for the ecosystem to survive, we must imitate financial services and products that now exist in our conventional, fiat currency world and make them available to crypto, as well as other emerging cryptocurrencies.

Tendency Prediction

The financial analogue of the crypto market is, of course, the stock market. The area of stock market forecasting has expanded over the last decades to optimize financial return and has lately expanded with the arrival of high-frequency, low-latency trading devices paired with sophisticated machine learning algorithms. As a result, it should be expected that trend is likely process will be repeated in the realm of cryptocurrency as the internet grows liquidity and more individuals develop an interest in effectively investing in the system. To accomplish so, we consider it essential to use machine learning technology to forecast the trend of cryptocurrencies.

Workflow

It requires historical data of any cryptocurrency in OHCL form to feed the machine learning algorithm. Data may be collected using either historical data or a real-time API [2]. We employed various machine learning algorithms with different parameters to train the model. We switched from candlestick chart to renko chart for a better outcome.

**Data Collection**

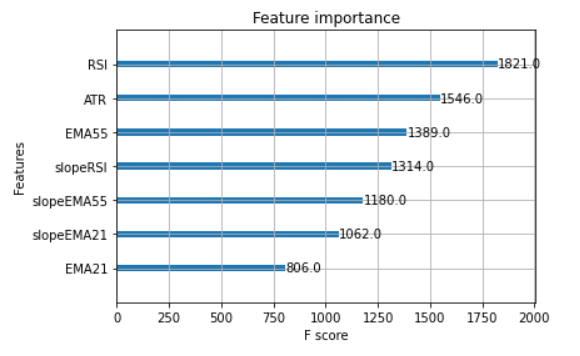
For our project, we gathered two sets of data. The first dataset contains 12600 minutes of data in OHCL (open, high, close, low) format. We utilize this information as ordinary work data. The second dataset contains more than 1.4 million minutes in OHCL (open, high, close, low) format spanning more than three years. We utilize this data for our final result or to verify the overall development of the model in reducing overfitting and improving forecast accuracy. We used bitcoin data to anticipate cryptocurrency trends because it has a high level of volatility. We obtained our data from a free web portal [dukascopy.com] The other is obtained using the online python WebSocket-client API.

**Feature Selection and Extraction**

We weighed in on more than 5 dependent elements of bitcoin price, market strength and technical indicators commonly utilized in trading. These attributes were chosen by hand based on our investigation into their relevance to the problem that we are attempting to address. After calculating and deriving these features using a third-party Python package, we add extra attributes that indicate the slope of all the features generated. For target variable into three distinct types such as 1: down-trend, 2: up-trend and 0: sideways-trend by considering distinct pip differences and prior candles. The targeted variable indicates whether the bitcoin market is moving up, down, or sideways at any given time.

Some of the features are determined based on the XGBoost Classifier to determine which feature variable influences how the model is trained.

According to the results, the most influential features for trend prediction are the RSI (Relative Strength Index), ATR (Average True Range), and various EMAs (Exponential Moving Averages) and their slopes.



**Why Decision Tree and Random Forest?**

A decision tree is a machine learning method that may be used to predict stock market trends. Since they can record complicated correlations between multiple variables and generate predictions based on these associations, decision trees are a popular choice for this purpose.

A decision tree model may be trained to anticipate stock trends using historical stock price data and other relevant variables such as trade volume, business financials, and economic indicators. Based on these characteristics, the model may then be used to predict the stock's future direction.

Ultimately, decision trees are simply one of several machine learning algorithms that may be used to anticipate market trends. The method used will be determined by the problem's precise needs, the available data, and the user's competence.

**Algorithm**

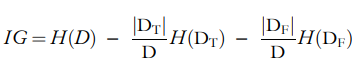
A common ensemble learning strategy is random forest, and decision tree is its poor learner. Based on the core concept, Random Forest may be classified as a bagging-integrated learning approach, which is an enlarged form of Bagging. Random forest is a mysterious black box methodology, while decision tree is a widely used white box classifier with visualization capabilities.

The process of creating the tree should conform to the simple notion, which indicates that the desired decision tree should be basic and compact. According to the Occam razor idea, the best algorithm is the most basic model that can interpret facts. The purity of a subset may be determined using information gained by Gini impurity. Take samples from either a set that corresponds to a hidden layer or a dendrogram, with the probability Pr(y) of y unique instance being proportionate to the set's fraction of the kind of instance. The Sampling distribution of labelled statistical distribution is used to assess the statistical uncertainties of the resultant class:

The entropy of such class statistical distribution serves to assess the chaotic level of such a set within information gain technique. The following formula encodes the units of information (entropy) into bit stream by using log2Pr (y).



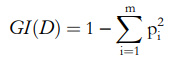
When all types of samples in such a set are evenly distributed, the entropy is greatest; when all cases belong to a single class, the entropy is lowest. The entropy shift at the start and end of a data set split signifies information gain. The instance set is represented by D, and D = DTUDT is a partitioning, and the information gain is characterized as follows:



Gini impurity is the frequency with randomly selected pieces from the collections are incorrectly labelled, as seen below:



where m represents the total number of classes and fi denotes the percentage of items in the ith collection. Using frequency to estimate probability, that is, fi ≈ pi from data set D, the research outcomes may be obtained:



In the machine learning field, the result of classification problems is often determined by a clear majority, but the result for regression problems is typically determined by an average value.

**Performing model and fitting**

Because this is time series data, it must be spitted in different manner rather than the train test spilt (linear\_model.train\_test\_spilt), with the first 80% of the data treated as training data and the remaining data used to test the model, analyze the classification model, and determine the model's efficiency.

Chart

Description automatically generatedChart

Description automatically generated

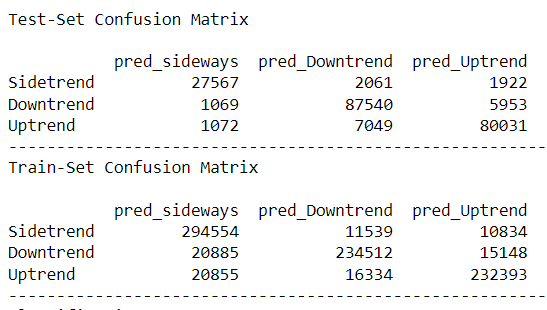
After splitting data into 80-20 manner and fitting it with both the classification algorithms (Random Forest Classifier and Decision Tree Classifier) the classification report for Random Forest is,

Table

Description automatically generated

In classification report, the precision, re-call and f1-score for all the classes as 0: sideways-trend, 1: down-trend and 2: up-trend. The accuracy of all the overall classification report is 91%.

For the performance of model in train data and test data, the confusion matrix of train data and test data is as per the given below that also defines the length of data and quality of model.



According to the confusion matrix, there are over 200000 samples of data for testing the trained model and over 830000 samples of data for training the model.

**Model Evolution**

For the evolution purpose of this model, some facts were manipulated. As of now we used a data of candlestick chart where condition of market can be in up-trend, down-trend or sideways-trend. Nevertheless, if we use the data of renko chart then there is no such thing as sideways trend. As a consequence, including renko chart data may increase the model's accuracy and yield better results than just candlestick data.

**What is Renko?**

In technical analysis, a renko chart is used to show how an asset's price changes. Given that the chart is comprised of bricks or blocks, it gets its name from the Japanese word "renga," which means "bricks".

Renko charts, as opposed to typical price charts, which display time-based bars or candles, are formed based on an asset's price movement and only generate a new brick when a specified price movement threshold is satisfied.

Renko charts can help traders filter out market noise and focus on price movements. They can assist traders in identifying crucial support and resistance levels and in detecting trend reversals.

**How to Convert Candlestick to Renko?**

1. Determine the brick size: The first step is to choose a brick size, which is the price at which a new brick will be added to the Renko chart. This can be determined by the volatility of the asset and the trader's preferences.
2. Calculate the Renko brick values: Beginning with the first candle, compare the candle's closing price to the closing price of the preceding Renko brick. If the difference between the two prices is larger than or equal to the brick size, a new brick is added to the Renko chart with the same opening price as the preceding brick's closing price, but a different closing price based on the price movement's direction.
3. Calculate the OHLC values for each Renko brick: Get the OHLC values for each brick after generating the Renko chart. The opening price of each block is the same as the closing price of the previous brick. The closing price is the cost of completing the brick. The high and low prices of the brick represent the highest and lowest price levels obtained during the brick's development.
4. Repeat the process for each subsequent candle: Continue comparing the closing price of each candle to the closing price of the preceding Renko brick and adding fresh bricks to the chart as needed.

It should be noted that Renko charts are entirely focused on price fluctuations and do not take time into account. As a result, the time intervals between bricks may differ, and the chart may not be appropriate for all trading techniques.

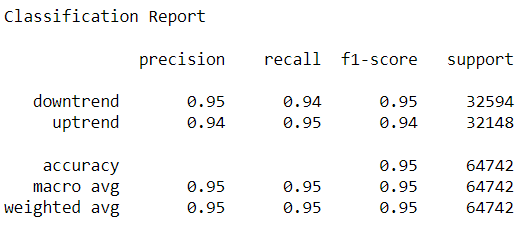
Chart, line chart

Description automatically generated

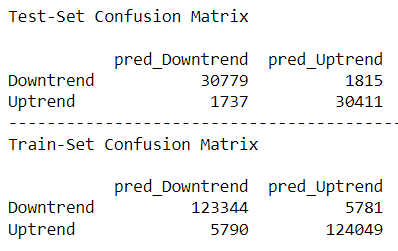


After changing the data from candlestick to renko, it must be in OHCL format in order to deduce the additional elements required to anticipate the trend, such as technical indicators and computing the target variable. To retrieve the target variable, we must repeat the process used to get it in candlestick data.

To run the machine learning algorithm, simply must repeat the steps that took in candlestick. After performing the random forest classifier, the classification report is,



The confusion matrix of train data and test data is as follows for the performance of the model in train data and test data, which also specifies the length of data and the quality of the model.



As per the confusion matrix, the data used to train the renko data is substantially less than the data used to train the candlestick data since we utilized renko brick size 3, which considers at least three candles as one brick. As a result, data must be reduced thrice. There is no sideways trend in renko data. As a result, data must be classified into only two classes, 1: down-trend and 2: up-trend.

**Result**

Here are the results of performing three different classification algorithms on data spanning more than three years, with intervals of one minute. We implemented the XGBoost, SVM, and Random Forest Classifier methods. The results of candlestick data are shown in the first table.

|  |  |  |  |
| --- | --- | --- | --- |
| Statistics | XGB Classifier | SVM | Random Forest |
| Sensitivity (TPR) | 0.3400 | 0.3333 | 0.9033 |
| Specificity (TNR) | 0.5395 | 0.0000 | 0.9239 |
| Precision (PPV) | 0.3366 | 0.1266 | 0.9166 |
| Accuracy (ACC) | 0.3400 | 0.3800 | 0.9100 |

The results of renko data are shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Statistics | XGB Classifier | SVM | Random Forest |
| Sensitivity (TPR) | 0.4950 | 0.5000 | 0.9550 |
| Specificity (TNR) | 0.7360 | 0.5548 | 0.9464 |
| Precision (PPV) | 0.4950 | 0.2750 | 0.9550 |
| Accuracy (ACC) | 0.5000 | 0.5500 | 0.9500 |

One can definitely observe a rise in the results, with accuracy increasing by more than 4% after changing from candlestick to renko data.

Conclusion

As it began with candlestick data from more than three years with intervals of one minute and more than one million samples, it was a data of more than one million minutes. We used machine learning algorithms to estimate the trend of the current market condition after deriving different feature variables and target variables. Using the Random Forest Classifier, we got a maximum accuracy of 91%. Data must be modified to renko pattern, which is asset volatility-based chart pattern, for further model evolution. By incorporating renko data into a machine learning model, the model's accuracy increased by more than 4%.

**Future work**

In order to enhance our achievements in the future, we propose to investigate multiple patterns of subsets of data, match them using RNN and apply them in parallel to more precisely determination of the current market condition. Because Long Short-Term Memory (LSTM) works on a short-term data memory and we do not require much additional knowledge about the market, we are eager to blend RNN and LSTM to obtain a more precise result.

**References**

1. Equations and Explanation of Algorithm- CAAI Trans on Intel Tech - 2021 - Yin - Research on stock trend prediction method based on optimized random forest (2021)
2. Sharma, N., Juneja, A.: LSboost is used to combine random forest estimates for stock market index prediction. In: 2017 2nd International Conference on Technological Convergence. 1199–1202. IEEE, Mumbai, (2017).
3. API Coinbase - https://www.coinbase.com/docs/api/overview
4. E.F. Fama, K.R. French: Typical risk factors in stock and bond returns (1993) J. Finacc. Econ., 33 (1993), pp. 3-56
5. E.F. Fama, K.R. French: An asset pricing model with five factors (2015) J. Finacc. Econ., 116 (1) (2015), pp. 1-22
6. M. Zhu, D. Philpotts, R. Sparks, M.J. Stevenson: A hybrid technique to stock ranking that combines CART and logistic regression J. Portfolio Manag., 38 (1) (2011), pp. 100-109
7. M. Zhu, D. Philpotts, R. Sparks, M.J. Stevenson: The advantages of using tree-based models for stock selection J. Asset Manag., 13 (6) (2012), pp. 437-448
8. Sharafati, A., Asadollah, S., Hosseinzadeh, M.: The promise of new ensemble machine learning models for predicting effluent quality indicators and the uncertainty associated with them. Process Saf. Environ. Prot. 140, 68–78 (2020)
9. Sharafati, A., et al.: Use of newly built ensemble machine learning models for daily suspended sediment load prediction and uncertainty analysis. Hydrol. Sci. Journal/Journal des Sci. Hydrol. (2020)
10. Misra, P., Chaurasia, S.: Predicting the direction of a stock index utilizing two stages of machine learning model hybridization. In: 7th International Conference on Reliability, Information Technology, and Optimization, 2018. (Trends and Future Directions). (ICRITO), pp. 533–537. IEEE, Noida, (2018)
11. Lili Yin, Benling Li, Peng Li, Rubo Zhang.” Research on stock trend prediction method based on optimized random forest", CAAI Transactions on Intelligence Technology, 2021
12. Zheng Tan, Ziqin Yan, Guangwei Zhu. "Stock selection with random forest: An exploitation of excess return in the Chinese stock market”, Heliyon, 2019