INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

An approach for Safeties and Cryptocurrency Trading Using Investigative Data Analysis and AI

[1]SHAIK CHINA HUSSAIN, [2]D. RAMMOHAN REDDY

[1]PG Scholar, Dept. of MCA, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract: Exploratory data analysis (EDA) on selected technical indicators is performed prior to modelling, and more realistic models are developed by introducing a new reward loss function that maximises profits during training in this paper's discussion of the use of AI in the trading of securities and cryptocurrencies. Back-testing on two securities using Artificial Neural Networks (ANN) and Random Forests (RF) as discriminative models compared to their counterpart Nave Bayes (GB) as a generative model confirmed that EDA's findings that discriminative classification models can better capture the complex patterns within the data were correct. The new reward loss function is used to retrain the ANN with testing on AAPL, IBM, BRENT CRUDE, and BTC using auto-trading strategy as the intelligent unit, and the outcomes show that this loss superiorly out performs the conventional cross entropy used in predictive models, indicating that this loss superiorly enhances the learning process. The findings of this study point to the need for increased emphasis on EDA and increased attention to practical losses in the study of machine learning models for stock market prediction applications.

Keywords: securities, cryptocurrency, stock market, artificial intelligence, machine learning, probabilistic modelling, classification models, artificial neural network, random forests, naïve bayes.

I. INTRODUCTION

Predicting the direction (up or down) of stock price movement in the following k time steps is the basic goal of stock market prediction. If maximising profits is the primary goal of the prediction, then the issue is better posed as a classification than as a regression. According to Icon's discussion in [1], there are two primary assumptions used in stock market forecasting. The first hypothesis, known as the "Efficient Market Hypothesis," states that the behaviour of stock prices is

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

random and unpredictable. On the other hand, another theory proposes that stock markets are predictable, at least to some extent, and that this predictability translates into long-term profitability. In reality, there are several publications in the scholarly literature that lend credence to the second possibility. In particular, studies are still being conducted to determine whether or not probabilistic categorization methods may be used to generate profits. The choice of a features set for training and the identification of models that are ideally suited to this set is one of the primary difficulties in stock market modelling for prediction purposes. The philosophical question of "features and model selection" revolves on the thought process that went into deciding which characteristics to use in the first place.

Selecting the right characteristics is critical not just for developing superior AI models but also for real-world applications like stock market forecasting. Expert Graphs and Charts

Traders and analysts may choose from a seemingly unlimited number of technical indicators to optimise their trading approach. Exploratory data analysis (EDA), which focuses on visualising the characteristics, should be undertaken after picking the features (technical indicators) in order to construct a trustworthy classification model. Kuhn [2] argues that before trying modelling, it is also crucial to understand the statistical distribution of the classes and the characteristics, failing which the models may result in poor performance. In addition to features and models selection, it is crucial to create and train practical models that learn data patterns associated with maximisation of returns as opposed to ones that optimise classification accuracy. To the best of our knowledge, there is surprisingly little active research into the use of supervised machine learning with modified loss function that match the objective of trading, despite the fact that works in the literature show the outcomes of their models based on profits.

This work's original contribution is a proposed technique for AI trading that takes a methodical approach to model development and training. Specifically, this entails formulating and testing two hypotheses, both of which assert that increasing returns requires either engaging in EDA or adding a new kind of reward loss.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

II. REVIEW OF LITERATURE

A. A Strategy for Choosing Appropriate Data and Models

There aren't many works that conducted data analysis in order to choose characteristics and models for stock market prediction. He's work in [4] is impressive because he discusses a wide range of complex features selection methods that can be applied to stock market prediction, including Genetic Algorithm (GA), Principal Component Analysis (PCA), and Sequential Features Selection (SFS); however, he didn't continue his work by testing their efficacy in the application (by, for example, applying them on models to test which one is best). Based on the findings of linearity testing, Basak [5] used the Gini Impurity methodology for features selection as a wrapper method, and the non-linear classifiers Random Forest and extreme Gradient Boost were applied.

After generating 84 features of technical indicators over many time horizons and using them to train a Support Vector Machine (SVM) model without first undergoing EDA, Di employed Random Forest to narrow down the features. Although it is well-known that Pearson Correlation Coefficient (PCC) is only beneficial for detecting linear relevance and is appropriate for numerical type data, Cheng in [7] employed PCC for supervised filtering of features, that is evaluating numerical to categorical data relevancy. Despite being an essential part of creating a classification model with strong performance, doing exploratory research on the data remains an obvious hole in the literature of stock market prediction.

B. Usefulness of AI Models,

When creating AI models for real-world applications like price prediction or trading, realism and usability are paramount. Since the focus of this study is on the trading side of things, the relevance of the models mentioned in the literature will be evaluated in that light. There is a large body of literature that uses metrics other than profitability to assess a model's efficacy when back-tested using an out-of-sample dataset. In general, this criterion for assessment is not useful in either predicting or trading. While Basak's suggested study in [5] did indicate a high degree of accuracy for a 90-day trading window (>90%), it was not possible to calculate potential earnings. Keep in mind

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

that overnight financing is applied by brokers when a stock is held for an extended length of time in the context of leveraged trading using Contracts for Difference (CFD) and short selling, as was anticipated in the study. Similarly, no evaluation of the models described in [6], [8], or [11] was based on the results that were created using data that was not previously visible.

It's worth noting that many published publications use financial success as a yardstick for success. However, this is seldom paired with a back testing environment that accurately represents the conditions of live trading.

Although Gerlein employed profits as a gauge and emphasised the significance of trading expenses like spread and slippage in [9], these costs were not included into profit estimates, leaving the model's applicability unproven. The work published by CarapuQo in [10] is similar in that it simply takes into account the spread cost and ignores any transaction costs.

Although Li's [12] return estimates include account transaction and spread fees, the transaction cost is small (0.05%) for the leverage trading used in the trading technique.

III. OBJECTIVE, HYPOTHESES DESCRIPTION AND Methodology

Objective The main objective of this work is to develop a probabilistic classification model for trading applications, i.e., the probability of the class, either the asset will go higher or lower in price at t+1, given the features (input) at t, and it is given in the following general form:

$$p(C_k|\boldsymbol{S}) = f_k(\boldsymbol{W}, \boldsymbol{S})$$

Where Ck is the class for k = 1, 2, and S is the space of features to be used in the training process, W is the set of parameters to be estimated, and f is the mapping function.

B. A Description of the Hypotheses

The first hypothesis is that more realistic and profitable models may be generated by first doing substantial numerical and exploratory (visual) data analysis on characteristics and classes.

Second Hypothesis: As long as profits are the primary concern, a loss function that maximises profits during training produces better returns than a loss that merely takes accuracy into account.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

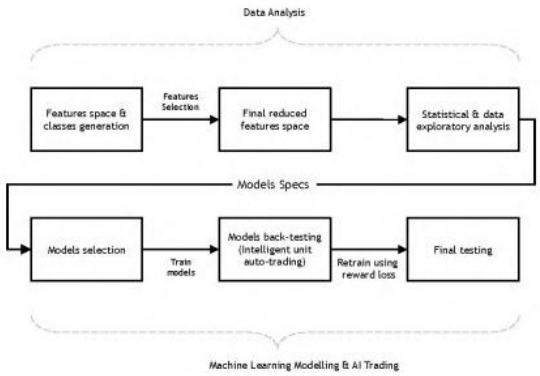

Fig. 1 Overview of the proposed methodology

TABLE 1 TIME PERIODS OF DATASET

Asset	Daily Prices Period	Hourly Prices Period
AAPL	Jan 2007 – Jan 2018	Jan 2017 – Dec 2020
IBM	Jan 2007 – Jan 2018	Jan 2017 – Dec 2020
OIL	--	Mar 2019 – Dec 2020
BTC	--	Mar 2019 – Dec 2020

TABLE 2 FINAL SELECTED FEATURES FORAAPL AND IBM

	AAPL	IBM
Indicator (i)	*Period (n)*	*Period (n)*
RSI	21, 22	8, 27
Fast %K	21, 49	12, 59
Slow %K	3, 5	35, 58
Slow %D	5, 19	34, 59
ADX	25, 43	10, 51
CCI	5, 18	3, 41
FRL	22, 49	12, 59

Methodology for Creating Models, Part C

The following phases (shown in Fig.1) are considered for the creation of the probabilistic models in order to achieve the goals of the work and to conduct efficient testing for the hypotheses:

First, get your hands on the raw information, including the asset's opening, closing, high, and low prices for the time period in question.

Extraction of characteristics (technical indicators) that have been created by humans.

Third, quantitative data analysis using features selection techniques.

4. Use exploratory data analysis (visualisation) to evaluate the connection between certain characteristics and classes, statistical distributions of data, and the data's dependence on the passage of time.

Fifth, suggest and train reasonable probabilistic models.

Testing the first hypothesis by evaluating the accuracy and profitability of the chosen trained models.

Update your stock prediction models by using a reward loss function that makes more sense in the real world.

Fig. 2 AAPL SU (left) and t-Stat (right) tests results

Reassessing the models with using the reward loss (test of second hypothesis) and building auto-trading strategy that serves as an intelligent trading unit (AI).

IV DATA ANALYSIS

A. Identifying and Choosing Relevant Features

1) Cost Estimates for Obtaining Raw Data

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

Share, commodity, and cryptocurrency price quotations will all be retrieved. These include the stock markets of Apple and IBM, the commodities market of Brent Crude oil, and the cryptocurrency market of Bitcoin. The prices are obtained not only daily but also hourly as shown in Table 1.

The prices at each time interval (open, close, low, and high) make up the raw data. Stock prices were gathered from finance.yahoo.com and www.dukascopy.com on a daily and intraday basis, respectively; past stock split price adjustments were also taken into account.

2) The Development of Technical Indicators and Goals

There are three types of trading models based on the kind of technical indicators they use, as described by Katz [13].

The terms "Break-Out model," "Moving average model," and "oscillator-based model" all refer to these three types of models. Oscillatory indicators like the Relative Strength Index (RSI), Commodity Channel Index (CCI), fast and slow Stochastics (%K and %D), and the Fibonacci Retracement Level (FRL) will form the basis of the models developed for this study. As a result, seven (7) indicators have been used. Multiple instances of each metric are produced, covering a range of time periods in the past. Time frames of 3 days to 60 days are common. Extracted from the price quotes from t down to t - n - 1, an indicator i for a past time period n is indicated by a column vector si>n of length N (number of samples), where i = 1, 2,7 represents a specific indicator and n = 3, 4, 5...., 60 represents the indicator period. Technical indicators, such as the Relative Strength Index (RSI), are calculated using the time period from t to t minus 2. A set of features in the subspace S" where S",$ c S" V n will be used to represent each indication throughout all of its periods. Two features from each subspace S will make up the final space S, increasing the total number of features in the final space to 14. Each digit from 1 to 14 represents a feature in the final space S. As will be explored in detail in the following two subsections, the two features from each S" will be chosen using statistical analysis that assesses the relevance of each feature in the subspace, S"$ c S" V n, with the classes. The fundamental benefit of this overarching technique is that it can be used to train models using just the most significant oscillatory technical indicators, yet

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

allowing for the inclusion of all of the indicators outlined by Katz [13].

To facilitate the study, it is important to note that the characteristics were discretized only for data reasons. This is how we discretized all sir$ c S" V n, i features:

$$s_{i,n} : 0 \le s_{i,n,t} \le 100 \rightarrow \{0, 1, 2, 3, \ldots, 100\}$$

For each observation of technical indicator i throughout time interval n at instant t, si>n>(is the corresponding expression. Because the Commodity Channel Index does not have a natural range of values between 0 and 100, it was rescaled before being discretized.

Target at time t for the dataset was specified as follows since it is known that the issue is a binary classification problem (either the price will rise or fall in the next time step).

$$T_t = \begin{cases} C_1 = 1; & close_{t+1} - close_t > 0 \\ C_2 = 0; & close_{t+1} - close_t \le 0 \end{cases}$$

3) Symmetric Uncertainty (SU) The SU between a feature and the target (two random variables) is the normalized mutual information between the two and thus can be given by:

$$SU(s_{i,n}, T) = 2 \times \frac{MI(s_{i,n}, T)}{H(s_{i,n}) + H(T)} \qquad (1)$$

Where MI and H are the mutual information and entropy respectively. 4) t-Statistics The usefulness of t-Stat method is that it calculates the normalized difference between the mean of the class conditional probabilities i.e., pJ si>n| C%K and pJ si>n\C*), and it is given by:

$$TS = \frac{\sqrt{N}\left(\overline{s_{i,n,C_1}} - \overline{s_{i,n,C_2}}\right)}{\sqrt{\sigma^2_{s_{i,n,C_1}} + \sigma^2_{s_{i,n,C_2}}}} \qquad (2)$$

The results of SU and t-stat tests for AAPL, daily data, are illustrated in the bar charts in Fig.2. For each indicator space S", only vectors S "n with the highest and lowest values are shown. The green bars are associated with max and the red are with min. Based on this result and for illustration purposes, the final selected features, for both AAPL and IBM daily data, are shown in Table 2.

V. CONCLUSION

This study provided a prediction approach that places more emphasis on evaluating the chosen data using EDA methods in order to use more suitable models and to improve their applicability by incorporating a new reward loss function during training. Two hypotheses about EDA were defined and tested, and a novel reward loss function was introduced to accomplish the goals of this study. The test results provided substantial support for the first hypothesis, showing that the intricacy

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

of the data was beyond the capabilities of simple models like NBC, and that ANN and RF provided the most accurate representation of the data. To go on with the second hypothesis test, we retrained the ANN using a loss function that maximises profits. Following extensive testing, it was shown that this loss significantly outperformed other common losses of predictive models, including the well-known cross-entropy.

Based on these findings, future work on machine learning models for trading applications may benefit from a greater emphasis on EDA and a greater emphasis on actual losses. Our current projects are investigating whether or not more sophisticated EDA methods for sequential data processing might provide better outcomes.

In addition, having a model that can forecast more than one time step ahead is being investigated as a means of improving the decision-making process (trading strategy).

The reward loss presented in this study, in conjunction with a regression ANN model that predicts many steps into the future, is intended to improve profitability.

REFERENCES

[1] O. Ican and T. B. Qelik, "Stock Market Prediction Performance of Neural Networks: A Literature Review," International Journal of Economics and Finance, 9(11), 100., 2017 doi:10.5539/ijef.v9n11p100

[2] M. Kuhn and K. Johnson, Applied Predictive Modelling, Michigan, USA: Springer, 2013, pp. 419 – 445

[3] C. M Bishop, Pattern Recognition and Machine Learning, New York, USA: Springer, 2006, pp. 114 - 227

[4] Y. He, K. Fat Aliyev, and L. Wang, "Feature selection for stock market analysis," International conference on neural information processing (pp. 737-744). Springer, Berlin, Heidelberg. 2013, November

[5] S. Basak, S. Kar, S. Saha, L. Khaidem, and S. R. Dey, "Predicting the direction of stock market prices using tree-based classifiers". The North American Journal of Economics and Finance, 47, 2019, pp.552 567.

[6] X. Di, "Stock trend prediction with technical indicators using SVM," Independent Work Report, Stanford Univ, 2014

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

[7] Afreen Bari, Dr. Prasadu Peddi. (2021). Review and Analysis Load Balancing Machine Learning Approach for Cloud Computing Environment.Annals of the Romanian Society for Cell Biology,25(2), 1189–1195.

[8] S. Dey, Y. Kumar, S. Saha, and S. Basak, "Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting" PESIT, Bengaluru, India, Working Paper, 2016

[9] E. A. Gerlein, M. McGinty, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach" Expert Systems with Applications, 54, 2016, pp.193-207.

[10] Prasadu Peddi (2021), “Deeper Image Segmentation using Lloyd’s Algorithm”, ZKGINTERNATIONAL, vol 5, issue 2, pp: 1-7.

[11] M. Qiu, and Y. Song, "Predicting the direction of stock market index movement using an optimized artificial neural network model," PloS one, 11(5), p.e0155133, 2016

[12] Y. Li, W. Zheng, and Z. Zheng, "Deep robust reinforcement learning for practical algorithmic trading," IEEE Access, 7, 2019, pp.108014 108022.

[13] J. O. Katz, and D. L. McCormick, The encyclopaedia of trading strategies. New York: McGraw-Hill, 2000, pp. 83 - 153

[14] Prasadu Peddi (2023), Using a Wide Range of Residuals Densely, a Deep Learning Approach to the Detection of Abnormal Driving Behaviour in Videos, ADVANCED INFORMATION TECHNOLOGY JOURNAL, ISSN 1879-8136, volume XV, issue II, pp 11-18.

[15] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction., Springer Science & Business Media, 2009