IMPERIAL

Innovative Approaches to Asset Prediction: Combining Deep Learning with Financial Modelling

Final Report

Author

C. Ioannidis

CID: 02533490

Supervised by

Prof. Pasquale Della Corte

Prof. Walter Distaso

Dr Lluis Guasch

A Thesis submitted in fulfillment of requirements for the degree of
Master of Science in Applied Computational Science and Engineering

Department of Earth Science and Engineering
Imperial College London
2024

Abstract

This project addresses the increasing complexity and volatility in financial markets through the development of advanced analytical tools. Leveraging the theoretical foundations established by Bryan Kelly and Kusuma, we propose refining Convolutional Neural Networks (CNNs) to enhance the prediction of financial asset behaviors.

Declaration of Originality

I hereby declare that the work presented in this thesis is my own unless otherwise stated. To the best of my knowledge the work is original and ideas developed in collaboration with others have been appropriately referenced.

Copyright Declaration

Contents

1
Introduction

Contents

1.1 Introduction

The prediction of financial markets has undergone significant evolution in recent years, driven largely by advancements in machine learning techniques. This study explores the use of Convolutional Neural Networks (CNNs) as advanced predictive models for analyzing time-series data across various asset classes. The primary objective of this research is to enhance the accuracy of financial market predictions by employing CNN architectures capable of capturing the complex patterns inherent in financial data. This project involves a comprehensive approach, beginning with the collection and preprocessing of extensive time-series datasets, followed by the development and iterative refinement of CNN-based models that address the unique challenges posed by financial market forecasting.

1

1.2 Background and Motivation

Traditional methods for predicting financial markets, such as the Auto Regressive Integrated Moving Average (ARIMA) model, have been widely applied but often struggle to capture the nonlinear and intricate dynamics of financial data. The advent of machine learning, and specifically CNNs, has introduced more sophisticated techniques capable of addressing these limitations. Recent research has demonstrated the efficacy of CNNs in financial market prediction by transforming time-series data into visual formats that can better capture underlying patterns.

For instance, Kusuma et al. (2019) utilized CNNs to analyze historical stock data converted into candlestick chart images, achieving high prediction accuracy for stock markets in Taiwan and Indonesia [1]. Sezer and Ozbayoglu (2019) further advanced this approach by transforming time-series stock data into 2-D bar chart images and applying CNNs to identify trading signals, demonstrating superior performance to traditional methods, especially during bearish market conditions [2]. Zeng et al. (2021) expanded on these concepts by introducing a video prediction model for economic time series that leveraged CNNs' ability to detect spatial patterns in image sequences, outperforming traditional techniques like ARIMA and Prophet [3]. Jiang (2023) also employed CNNs with OHLC (Open, High, Low, Close) charts to forecast stock returns, highlighting the model's capacity to recognize intricate patterns and adapt across different geographical and temporal scales [4]. Collectively, these studies underscore the potential of CNNs to significantly improve asset price prediction accuracy, offering a marked advantage over traditional forecasting methods.

1.3 Research Objectives

The primary goal of this research is to enhance the predictive capabilities of financial market models through the development and refinement of CNN architectures. By focusing on the analysis of time-series data, this study aims to uncover complex patterns that drive market behavior, thus facilitating more accurate and reliable predictions. Achieving higher predictive accuracy is expected to support more effective risk management strategies, particularly in volatile market environments. Additionally, this research seeks to provide insights that can assist investors and financial analysts in making more informed decisions regarding investment strategies, thereby contributing valuable knowledge to the field of financial analytics. The overarching aim is to advance the adoption of more sophisticated machine learning techniques in financial forecasting, promoting a deeper understanding of global market dynamics.

1.4 Methodology Overview

The methodological approach of this study encompasses several stages. Initially, extensive time-series data will be collected and preprocessed to ensure it is suitable for deep learning applications. The data will then be transformed into image formats, such as candlestick charts, to facilitate the training of CNN models. Following data preparation, various CNN architectures will be developed and iteratively refined to optimize their predictive performance. The models will be trained using historical market data to learn patterns and trends that may inform future market movements.

In the final stages of the project, the developed models will be tested within simulated environments to assess their accuracy and practical applicability in real-world scenarios. This evaluation will be conducted through rigorous backtesting and benchmarking against standard market indices, such as the S&P 500, to measure their performance relative to traditional investment strategies.

1.5 Expected Contributions

The anticipated outcomes of this research are expected to provide significant insights into market dynamics, thereby enhancing the decision-making processes in financial investments. By demonstrating the practical utility of CNNs in financial forecasting, this study aims to bridge the gap between theoretical advancements and their real-world applications in financial markets. The findings are expected to contribute substantially to the field of financial analytics, promoting the integration of advanced machine learning techniques into market prediction and portfolio management strategies. This research represents a meaningful contribution to the ongoing discourse on the application of deep learning in financial markets, with implications for both academic research and practical investment decision-making.

2
Methodology

Contents

2.1 Training

2.1.1 Data Preprocessing and Integration

The preprocessing phase was a critical step in preparing the data for effective training of the Convolutional Neural Network (CNN) model. Given the nature of financial time-series data and the need to convert it into formats suitable for deep learning, a rigorous and systematic approach was adopted.

Initially, the data underwent a comprehensive cleaning process to address inconsistencies, missing values, and outliers. As the data was sourced from multiple repositories—namely CRSP, Kaggle, and Yahoo! Finance—standardization across datasets was paramount. Missing values were handled using techniques such as forward and backward filling to maintain the temporal continuity of the data, while outliers were identified and removed using statistical methods like Z-score analysis and interquartile range (IQR) filtering. This step was essential to ensure that the data fed into the CNN models was both reliable and representative of typical market behaviors.

Following the cleaning process, normalization was applied to standardize the range of the OHLC (Open, High, Low, Close) data. Normalization is crucial in neural network models to ensure that all input features contribute equally to the learning process. This typically involved scaling the OHLC data to a range between 0 and 1, which helps in achieving faster convergence during the training phase and avoids biases that could arise from differing magnitudes in raw data values.

A significant aspect of the preprocessing phase was the transformation of the normalized OHLC data into 64x64 pixel images, specifically candlestick charts, which serve as the primary input format for the CNN models. The transformation process involved converting sequential OHLC data points into a series of candlestick charts, capturing the temporal dynamics and price movements over fixed intervals. These images were then stored in .npy files, an efficient format for handling large-scale image datasets within NumPy arrays, enabling streamlined data loading and manipulation during the training phase.

This approach of converting OHLC time-series data into visual representations leverages the ability of CNNs to recognize complex spatial patterns, which are indicative of underlying market trends and behaviors. By using a visual representation, the CNN model is better positioned to learn from patterns in the data that are not easily captured through traditional numerical methods, thus enhancing the robustness and accuracy of the predictive model.

2.1.2 Model Development and Training

The development and training of the CNN model were meticulously structured to explore various architectural configurations aimed at maximizing predictive accuracy while minimizing overfitting. The chosen architecture treated the model as a classifier, designed to predict market direction based on the input images derived from OHLC data.

The initial phase of model development involved experimentation with multiple CNN architectures to identify the most effective structure for financial market prediction. This experimentation included testing various combinations of convolutional layers, pooling layers, and activation functions. Key architectural features such as dropout layers were integrated to reduce overfitting by randomly deactivating a subset of neurons during each training iteration. This technique helps to generalize the model, ensuring it does not overly fit the training data at the expense of performance on unseen data.

Residual blocks were also employed to enhance the model's depth and ability to capture complex features. Residual connections facilitate the training of deep networks by allowing gradients to flow more effectively through multiple layers, preventing the vanishing gradient problem commonly encountered in deep learning. This capability is particularly valuable in financial modeling, where deep architectures can uncover complex, non-linear relationships inherent in market data.

In addition to standard CNN layers, the model architecture incorporated Long Short-Term Memory (LSTM) layers to capture sequential dependencies in the time-series data. LSTMs, a type of recurrent neural network (RNN), are well-suited for handling sequences and can retain information across time steps, making them ideal for capturing the temporal dependencies crucial in financial forecasting.

The training dataset, encompassing randomly selected stocks from the U.S. market from the 1990s until 2017, was deliberately chosen to ensure that the model trained on data entirely unrelated to the backtesting period (June 2019 to June 2024). This strategic choice was aimed at preventing any data leakage between the training and testing phases, thereby enhancing the model's generalizability and ensuring that performance metrics reflect true predictive power rather than overfitting to historical data.

The training process involved optimizing the model's parameters through iterative tuning of hyperparameters such as learning rates, batch sizes, and epochs. Optimization algorithms, including Adam and RMSprop, were explored to identify the most effective approach for minimizing the loss function and achieving stable convergence. The model's performance was continuously evaluated using classification metrics such as accuracy, precision, recall, and F1 score, which provided an outline of which configurations were likely to perform best during the subsequent backtesting phase.

By rigorously testing and refining these CNN architectures, the study aimed to identify the model configuration that most effectively captures the complex patterns in financial data, providing a robust tool for predicting market movements and contributing valuable insights to the field of financial analytics.

2.2 Evaluating

2.2.1 Model Evaluation and Backtesting

The evaluation of the Convolutional Neural Network (CNN) model was conducted through a two-step process that involved both classification performance assessment and a comprehensive backtesting strategy. The aim was to assess the model's effectiveness in predicting market movements and to evaluate its practical applicability within a real-world trading environment.

Model Evaluation

The CNN model, designed as a classifier, was initially evaluated using standard classification metrics: accuracy, precision, recall, and F1 score. These metrics provided a comprehensive overview of the model's performance in distinguishing between various market conditions, such as bullish, bearish, or neutral trends.

- **Accuracy** was utilized to measure the proportion of correct predictions out of the total number of predictions, offering a broad sense of the model's overall performance.
- **Precision** was calculated to determine the proportion of true positive predictions out of all positive predictions made by the model, indicating how effectively the

model identifies market conditions that it forecasts to occur.

- **Recall** was assessed to understand the model's sensitivity in correctly identifying all actual instances of a specific market condition.

- **F1 score**, the harmonic mean of precision and recall, was used to balance the trade-off between these two metrics, particularly in financial market prediction contexts where false positives and false negatives have different implications on trading strategies.

These evaluation metrics provided a detailed assessment of the CNN models likely to perform well in practical trading scenarios. Based on these evaluations, the models that demonstrated the best balance between accuracy, precision, recall, and F1 score were selected for the subsequent phase: backtesting.

Backtesting Strategy

Following the initial model evaluation, the selected CNN models were subjected to rigorous backtesting to assess their performance within a simulated trading environment. The backtesting covered a period from June 2019 to June 2024, using historical market data to evaluate the models' predictions and trading decisions.

The backtesting strategy employed a weekly rebalancing approach, consistent with the lookback periods used during the model training phase. This strategy, inspired by the methodology detailed in Kelly's and Jiangs's paper [4], ensures coherence between the training and testing phases, thereby providing a robust framework for evaluating the model's predictive capabilities over time. The weekly rebalancing strategy allows the model to adjust its positions based on updated signals, reflecting a dynamic trading approach that is responsive to evolving market conditions.

The CNN model generated signals indicating whether to go long, short, or hold a position based on its predictions. A "long" signal indicated a positive market outlook, prompting the strategy to purchase and hold the asset, while a "short" signal suggested a negative outlook, prompting the strategy to sell or short the asset. A "hold" signal indicated a neutral outlook, suggesting no changes to the current position. This tripartite strategy is designed to capture market opportunities while managing downside risk, leveraging the CNN model's predictive power to guide trading decisions.

The backtesting was conducted using QSTrader, an open-source framework for implementing systematic trading strategies. QSTrader enabled a comprehensive analysis of the strategy's performance relative to a buy-and-hold strategy of the S&P 500 index over the same period. The buy-and-hold strategy served as a benchmark, representing a passive investment approach commonly employed in the market.

To evaluate the effectiveness of the CNN-driven trading strategy, several key performance metrics were analyzed:

- **Cumulative Return**: The total return of the portfolio over the backtesting period, providing a direct comparison between the growth of the portfolio under the CNN strategy and the buy-and-hold strategy.
- **Sharpe Ratio**: A measure of risk-adjusted return, calculated as the ratio of the portfolio's excess return over the risk-free rate to its standard deviation. This metric was used to assess how well the CNN-driven strategy compensated for risk relative to the benchmark.
- **Maximum Drawdown**: The maximum observed loss from a peak to a trough of the portfolio, before a new peak is attained. This metric was crucial for evaluating the risk exposure of the CNN strategy compared to the buy-and-hold benchmark.
- **Volatility**: The standard deviation of the portfolio's returns, indicating the level of risk or uncertainty associated with the strategy's performance.

The backtesting results demonstrated that the CNN-driven strategy, with its dynamic weekly rebalancing and responsiveness to market signals, outperformed the buy-and-hold strategy of the S&P 500 in terms of cumulative return and risk-adjusted performance. The strategy successfully captured significant market trends, both upward and downward, by dynamically adjusting its positions based on the CNN model's signals. The CNN strategy exhibited a higher Sharpe ratio, indicating superior risk-adjusted returns compared to the benchmark. Additionally, the maximum drawdown of the CNN strategy was lower than that of the buy-and-hold approach, suggesting that the model was effective in mitigating downside risk during market downturns.

By benchmarking the CNN-driven strategy against a buy-and-hold approach, the study effectively highlighted the potential advantages of utilizing advanced machine learning techniques for financial market prediction and portfolio management. The results underscore the practical utility of CNN models in developing systematic trading strategies that are not only predictive but also adaptable to varying market conditions. This research contributes to the expanding body of literature on the application of deep learning in financial markets, offering insights into how such models can be leveraged to enhance portfolio performance and manage investment risk effectively.

3
Results

4 Discussion

5
Conclussion

Bibliography

[1] R. M. I. Kusuma, T.-T. Ho, W.-C. Kao, Y.-Y. Ou, and K.-L. Hua, *Using deep learning neural networks and candlestick chart representation to predict stock market*, 2019. arXiv: `1903.12258 [q-fin.GN]`.

[2] O. B. Sezer and A. M. Ozbayoglu, *Financial trading model with stock bar chart image time series with deep convolutional neural networks*, 2019. arXiv: `1903.04610 [cs.LG]`.

[3] Z. Zeng, T. Balch, and M. Veloso, *Deep video prediction for time series forecasting*, 2021. arXiv: `2102.12061 [cs.CV]`.

[4] J. Jiang, B. Kelly, and D. Xiu, "(re-)imag(in)ing price trends," *The Journal of Finance*, vol. 78, no. 6, pp. 3193–3249, 2023. DOI: `10.1111/jofi.13268`. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13268`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13268`.