Causal Discovery in Stock Return

Xinqi Huang xih037@ucsd.edu

Ruizhe Dai rdai@ucsd.edu

Vivian Zhao vxzhao@ucsd.edu

Yishan Cai yic075@ucsd.edu

Mentor: Biwei Huang bih007@ucsd.edu

Mentor: Jelena Bradic jbradic@ucsd.edu

UC San Diego™

HALICIOĞLU DATA SCIENCE INSTITUTE

Introduction

- Driven by a complex interplay of economic conditions, corporate performance shocks, and investor anticipations, the high volatility of financial markets challenges stock return forecasting, further amplified by recent global policies and economic disruptions. As a result, capturing both long-term trends and external shocks has become crucial in guiding the decisions of millions of investors.
- Traditional forecasting models struggle with sudden market shifts due to their reliance on historical patterns and predefined economic relationships. Meanwhile, despite better accuracy, deep learning-based models lack interpretability, which limits their adoption in high-stakes financial markets.

To enhance model transparency and reliability, we proposed a hybrid stock return prediction framework that 1) uses PCMCI with DeepAR for causal feature selection and lag optimization to better capture general trends and 2) leverages CD-NOD on real-time macro-economic and company-level factors with nowcasting to capture the effect of external shock.

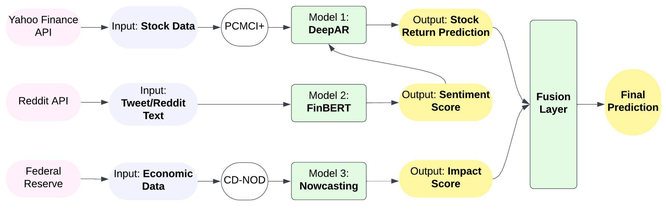

Fig. 1: Overview of the Proposed Stock Return Prediction Framework

Sentiment Analysis Module

FinBERT Sentiment Analysis Model Sentiment influences are important and influential to the performance of the stock market. FinBERT [1] is a pre-trained NLP model to analyze sentiment of financial text. FinBERT is a fine-tuned version of the BERT model and trained with many text under the financial environment, and it is suitable to perform financial sentiment classification.

Sentiment Data Collection We have found a dataset on GitHub repository that contains tweets about famous companies including: Apple, CVS, Ebay, dating from 2020-6-01 to 2023-5-31. However, our goal is to make our model more applicable, which we need more recent tweets to make our model better capture the pattern in the recent stock market. Initially we thought about scraping more tweets, but the cost of accessing X API was too costly for our group. Instead, we decided to use Reddit posts and comments, as Reddit's API is free. One obstacle is that Reedit's posts and comments are sporadic, and there are cases that in some day, there were no posts or comments discussing stocks of companies we want to scrape. Therefore, after using FinBERT to convert raw text regarding each company's text to numerical sentiment scores, we decided to use a linear interpolation method that is scaled by time differences to alleviate the irregular spacing of our sentiment data scraped from Reddit and also making the interpolation of our missing data more realistic.

Sentiment Score Calculation Using FinBERT's sentiment labels and their associated confidence levels, we computed our final sentiment score using the weighted sum of the sentiment labels (positive, negative, and neutral), where each label was weighted by its corresponding confidence level. These sentiment scores were then normalized to a range between 0 and 1 with an approximately

Stock Return Prediction Module

Data Collection & Processing Historical data of 6 companies (3 Tech, 3 Healthcare) from Dec 2019 to Jan 2025, including daily metrics of opening/closing price, high, low, and volume. We calculate Daily return as: $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$.

Feature Selection Applied PCMCI+ algorithm for causal feature selection, identifying 8 key covariates across temporal, price-based, and technical indicators based on their causal impacts.

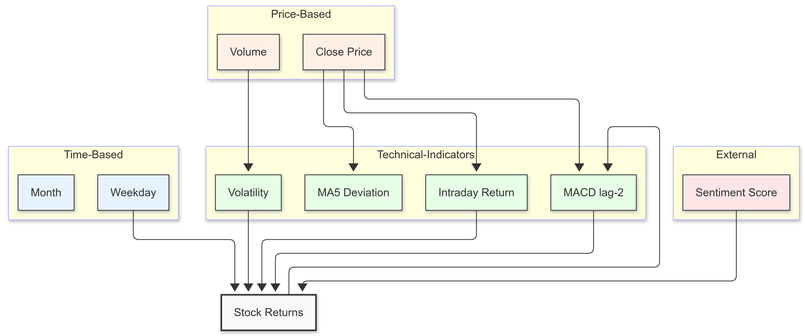

Fig. 2: Feature Relationship Diagram

Model Architecture Enhanced DeepAR model with four components: (1) Input layer combining historical returns z_{t-1}, technical indicators x_t, and entity embeddings e_i; (2) LSTM core with skip connections, orthogonal initialization, and variational dropout; (3) Attention mechanism for dynamic pattern weighting; (4) Probabilistic output layer producing return distributions $z_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$.

Economic Impact Analysis Module

Data Extraction We extracted microeconomic data from Yahoo Finance, including treasury shares, net debt, and sales of business, alongside macroeconomic indicators from FRED, such as Money Supply (M1, M2), Interest Rates, CPI, GDP, and Unemployment Rate. These datasets provide a comprehensive view of economic and firm-specific drivers of stock returns.

Frequency Alignment A key challenge was aligning data frequencies—GDP and financial reports are quarterly, while macroeconomic data is mostly monthly. We mapped each quarter to three monthly time-series vectors, applying regression analysis to determine the most representative months for quarterly stock return predictions.

Feature Selection Using causal learning (CD-NOD algorithm), we identified the most relevant predictors, reducing redundancy and enhancing model interpretability.

Nowcasting Forecasting The final nowcasting model integrates macroeconomic and firm-level financial data to predict stock returns in real time. This approach captures the dynamic interaction between economic trends and company performance, improving forecasting accuracy.

Results

We evaluate model performance using ND, RMSE, and Quantile Loss. Figure 3 shows prediction results for AMZN stock returns with 95% confidence intervals. Our model achieves ND/RMSE of 0.665/0.972 and 0.419/0.542 in two test periods respectively.

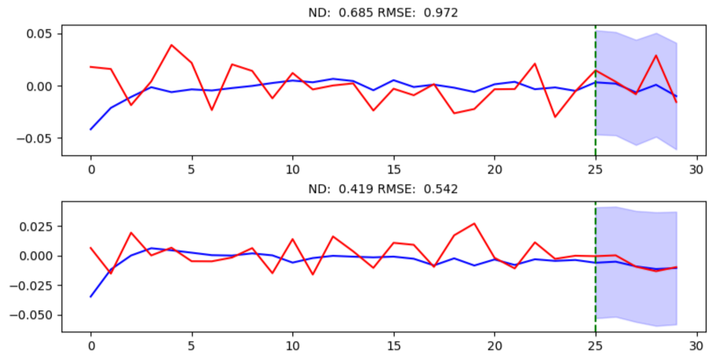

Fig. 3: AMZN Stock Return Prediction Performance

Remarks (Developing)

.....

Acknowledgements (Developing

We thank our mentors, Biwei Huang and Jelena Bradic, for their expert guidance and insightful suggestions on causal discovery and deep learning techniques. We are grateful to the University of California, San Diego for providing the opportunity to work on this meaningful project.

References(Developing)

[1] D Araci. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models". In: *arXiv preprint arXiv:1908.10063* (2019).