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 daily trends and 2) leverages **CD-NOD** on real-**time macro-economic** and **company-level factors** with Random Forest to capture the effect of shock on stock price.

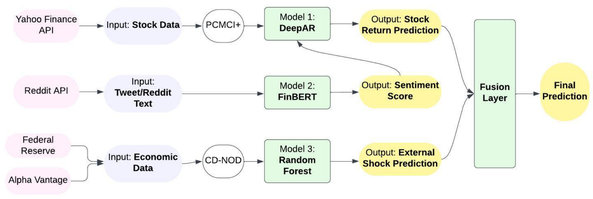
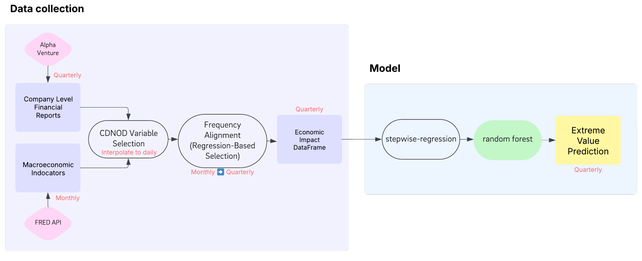


Fig. 1: Overview of the Proposed Stock Return Prediction Framework

Economic Impact Analysis Module



Data Collection

Microeconomic Data: Company quarterly reports from Alpha Vantage, including information on income statements, cash flow, and balance sheets

Macroeconomic Indicators: Monthly Economic data from FRED (CPI, GDP).

Data Pre-processing

Frequency Alignment: We first mapped each quarter to three monthly time-series vectors and then **regressed the monthly vector on the quarterly compound return**. Finally, we selecting the most representative month to aligned with company-level quarterly report via **hypothesis testing on the coefficients**.

Prediction Model:

A **Random Forest** model was applied to the integrated dataset to predict **extreme quarterly stock returns upon the release of company financial statements** in real time, to capture shocks driven by economic fluctuations and expectations.

CDNOD: Feature Selection from a Causal Len

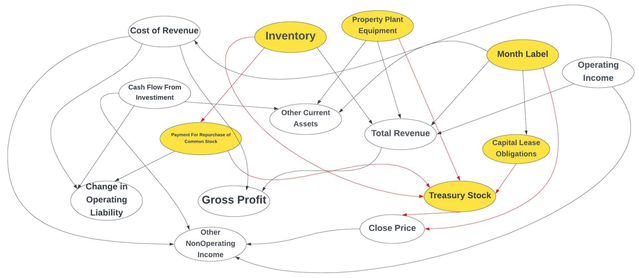


Fig. 3: Causal Graph for AT&T

Causal Feature Selection: After interpolating economic factors with daily stock prices, we applied **CD-NOD with monthly grouping** and **Fisher's Z-test at a 0.01 significance level** to capture causal relationships between factors and stock price shocks.

We define **impactful features** as those that (1) have a **direct edge** to stock price or (2) connect to stock price **through causal pathways** in the learned graph.

We further performed **pairwise regression** to assess each predictor's direct impact.

Stock Return Prediction Module

Data Collection & Pre-processing: Historical data of 6 companies (3 Tech, 3 Healthcare) from Jun 2020 to Feb 2025, including daily metrics of opening/closing price, high, low, and volume. We calculate Daily return as: $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$.

Causal Feature Selection: Applied PCMCI+ algorithm for causal feature selection, identifying 8 key covariates based on their causal impacts.

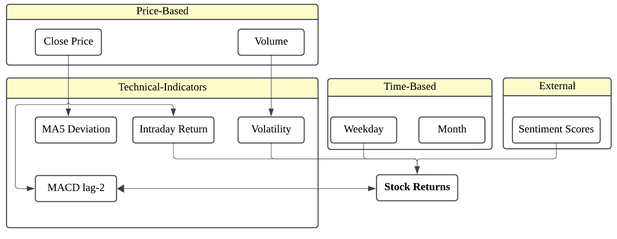


Fig. 4: Causal Structure of Stock Return Predictors after PCMCI Analysis

Model Architecture: The model integrates historical returns, technical indicators, and entity embeddings into a concatenated input tensor. An enhanced LSTM with skip connections and variational dropout enables robust gradient flow. The probabilistic output layer generates a Gaussian distribution of future returns, instead of just point estimates.

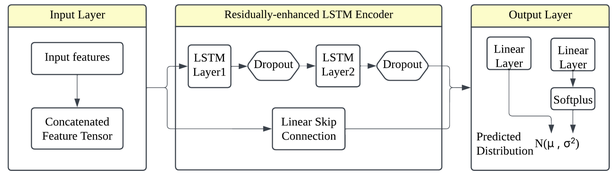


Fig. 5: DeepAR Model Pipeline

Sentiment Analysis Module

Model: FinBERT - Specialized NLP model for financial text sentiment analysis.
Data Collection & Pre-processing:

- AMZN, GOOG, CVS GitHub Tweet Dataset (June 2020 - May 2023).
- Collected Reddit posts and comments for additional sentiment data via API.
- Applied time-scaled linear interpolation for data smoothing.

Sentiment Scoring: Weighted FinBERT confidence levels with normalization to 0-1 range.

Fig. 6: Work Flow of Sentiment Analysis Module

Results

Evaluation Metrics: MAE, MAPE, RMSE, and Direction Accuracy.
In figure below, our model accurately captures overall market trends, with CVS stock price direction being predicted with 100% accuracy, while the AMZN model excels particularly in long-term forecasting with a long-term MAE of only $2.61.

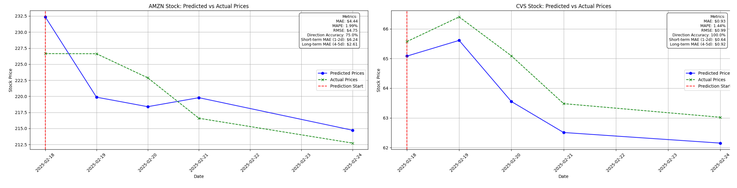


Fig. 7: Stock Price Prediction Performances

Remarks (Developing)

.....

Acknowledgements

We thank our mentors, Biwei Huang and Jelena Bradic, for their expert guidance and insightful suggestions on causal discovery and deep learning techniques.

References(Developing)