Causal Discovery in Stock Return

— Ensemble Deep Learning for Stock Return Prediction in Volatile Markets

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

- Financial market volatility is driven by a complex interplay of **economic conditions**, **corporate shocks**, **investor anticipations**, **global policies**, and **economic disruptions**, making stock return forecasting challenging.
- Capturing **long-term trends** and **external shocks** is crucial for informed investment decisions.
- Traditional models struggle with **sudden market shifts** due to reliance on historical patterns.
- Deep learning models, while more accurate, **lack interpretability**, limiting their adoption in finance.
- **Company List:** Amazon(AMZN), Google(GOOG), AT&T(T), Abbott Laboratories(ABT), Amgen(AMGN), CVS Health Corporation(CVS)

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.

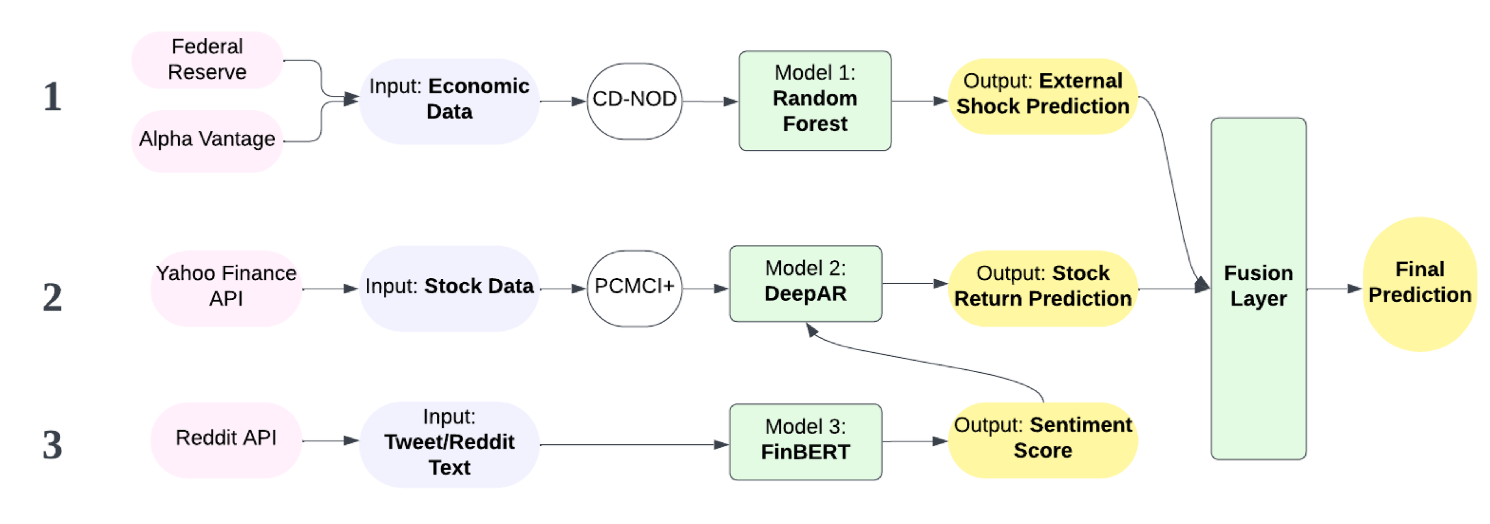

Fig. 1: Overview of the Proposed Stock Return Prediction Framework

1. Economic Impact Analysis Module

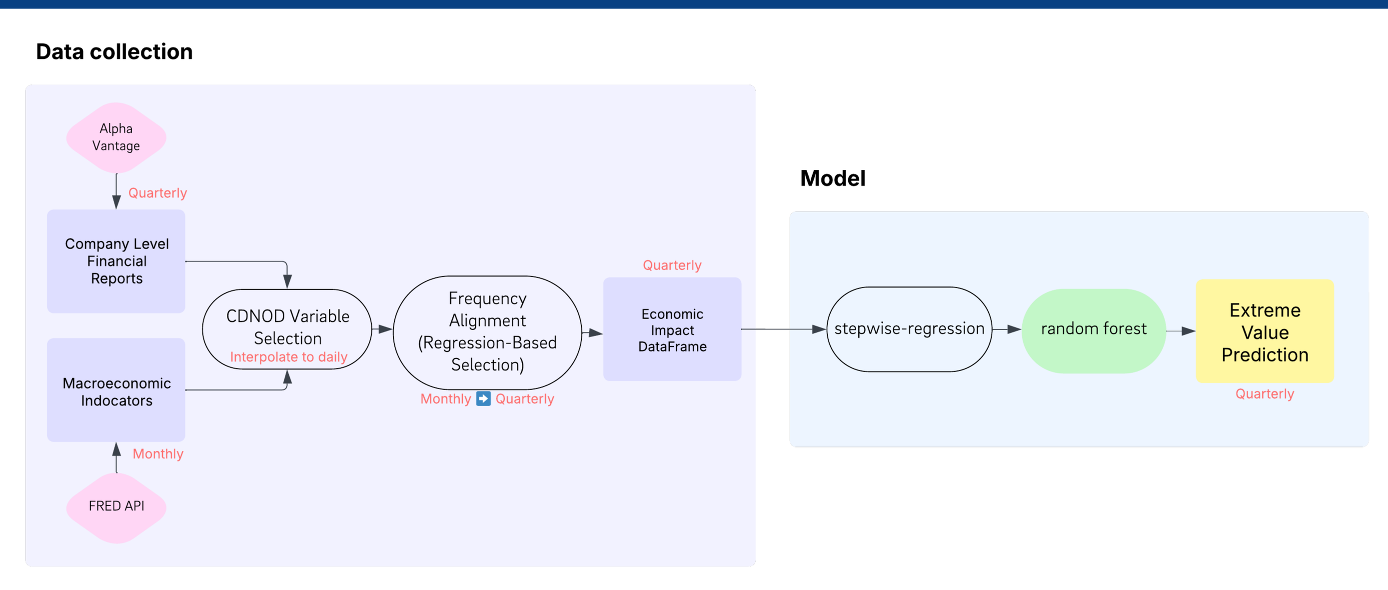

Fig. 2: Economic Impact Framework

Data Collection

- **Microeconomic Data** Company quarter reports (balance sheet, cash flow).
- **Macroeconomic Indicators** Monthly Economic data (CPI, GDP).

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

Prediction Model: **Random Forest** predicts **extreme quarterly stock returns upon the release of company financial statements**, to capture shocks driven by economic fluctuations and expectations.

CDNOD: Feature Selection from a Causal Lens

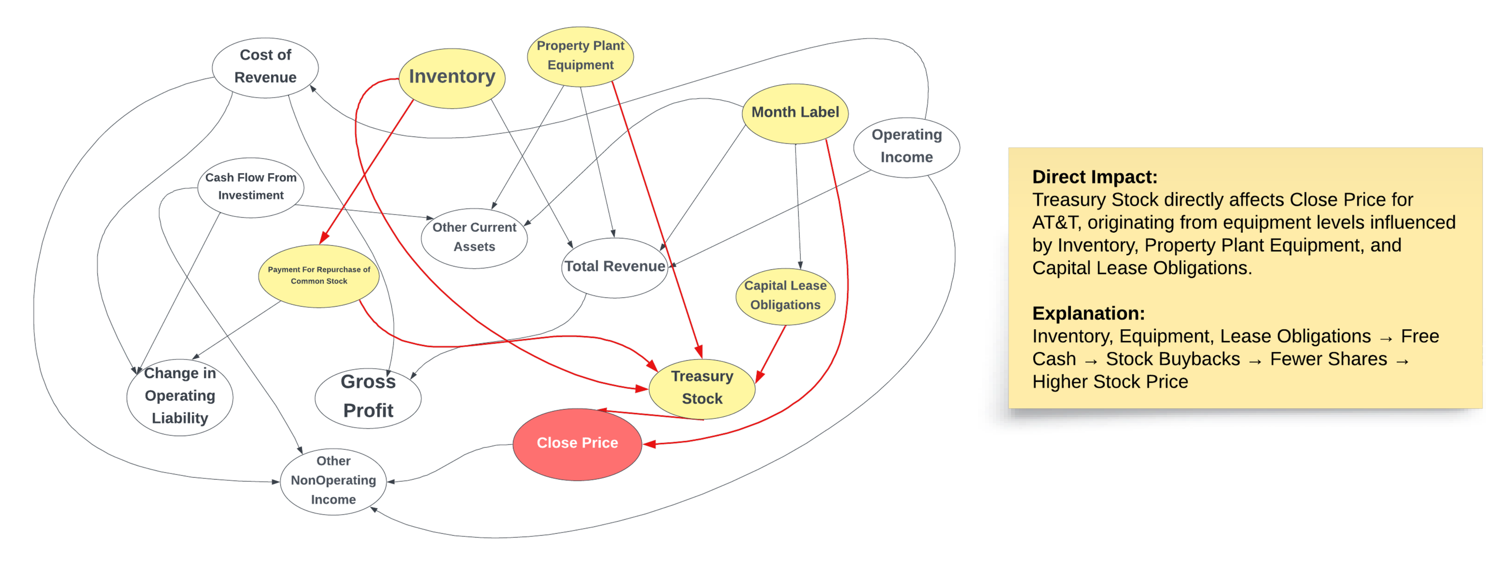

Fig. 3: Causal Graph for AT&T

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.

Defining impactful features

- have a **direct edge** to stock price.
- connect to stock price through **causal pathways** in the learned graph.

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

2. Stock Return Prediction Module

Data Collection & Pre-processing: We use 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.

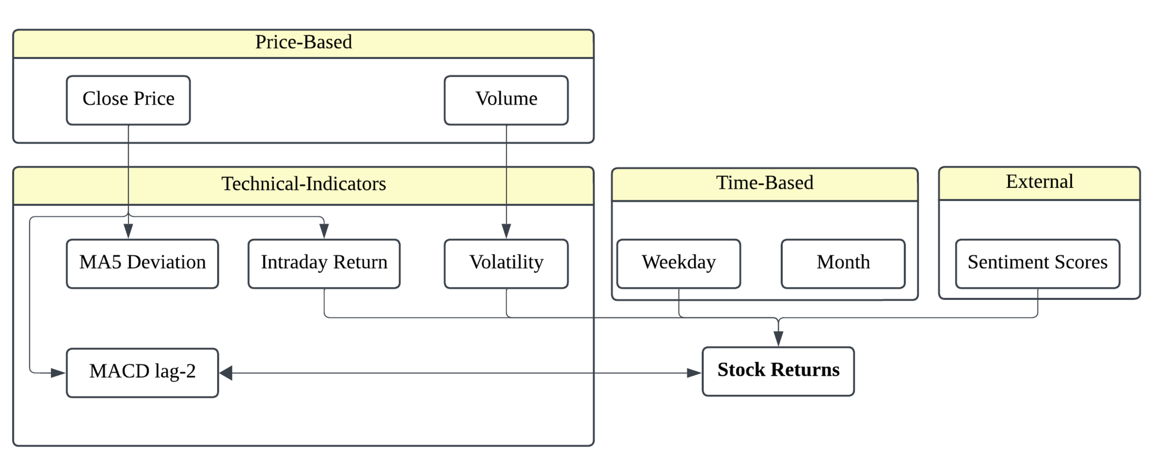

Fig. 4: Causal Structure of Stock Return Predictors after PCMCI+ Analysis

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.

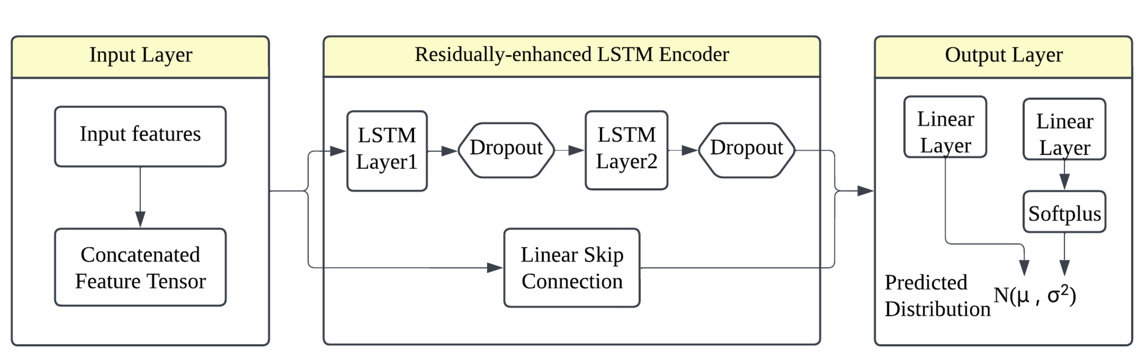

Fig. 5: DeepAR Model Pipeline

3. 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

Fusion Layer and Final Results

Fusion Layer architecture: This integrates DeepAR daily predictions with quarterly financial data through an adaptive weighting mechanism, improving stock price forecasting accuracy.

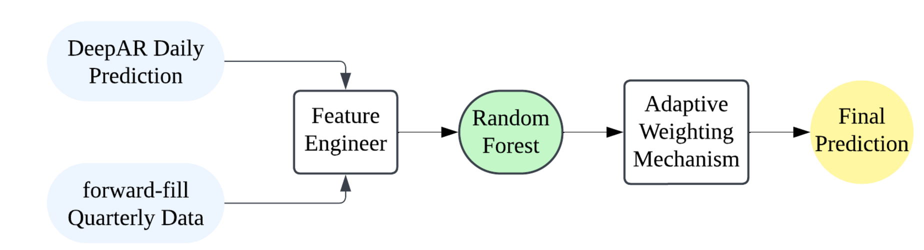

Fig. 7: Fusion Layer Architecture

Evaluation Metrics: MAE, MAPE, RMSE, and Direction Accuracy.
Our model captures market trends effectively, with ABT stock showing 100% direction accuracy and AT&T achieving exceptional short-term precision (MAE: $0.57).

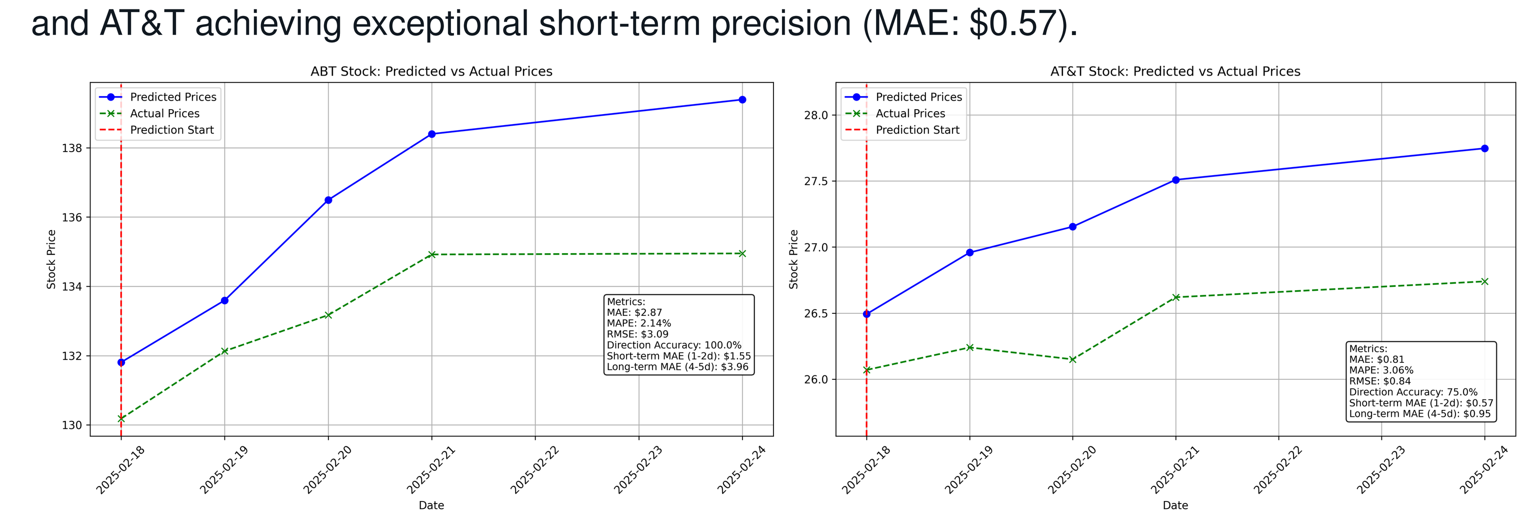

Fig. 8: Stock Price Prediction Performances

Acknowledgements

We sincerely thank our mentors, Biwei Huang and Jelena Bradic, for their guidance on causal discovery and our overall framework. We also appreciate Professor James Hamilton's advice on handling the frequency mismatches in the economic factors.

References

1. Araci, D. (2019). "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." arXiv:1908.10063.
2. Zhang, Y.; Jiang, Q. et al. (2019). "You May Not Need Order in Time Series Forecasting." arXiv:1910.09620.
3. Salinas, D.; Flunkert, V.; Gasthaus, J. (2017). "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks." arXiv:1704.04110.