A Comparative Analysis of Machine Learning Models for Cryptocurrency Return Prediction

Ke Su(Scott)

Internship Final Presentation

Thira Lab

THIRA

Part 1:Introduction and settings

Project Background & Objectives

- Objective
  - To systematically backtest and compare a diverse set of machine learning models (from XGBoost to CNNs) to identify the most profitable and robust architecture for predicting short-term cryptocurrency returns.

- My Role
  - Develop a Standardized Training and Backtesting Framework to ensure a fair and robust comparison across all tested architectures.
  - Evaluate Models: Systematically test and improve the performance of different type models.

Methodology: Data, Features & Target

- Asset Universe: the three most liquid crypto currencies (BTC, ETH, and BNB).
  - Data for these three assets is aggregated to train a single, cross-asset global model.
  - The data includes transaction cost data, hourly closing prices, and initial feature data.
  - The initial features consist of 156 initial features for each asset.

- Target Value Transformation:
  - The log return with a 6-hour forecast (log(P_t+6 / P_t)) (after experimenting with ranges from 1 to 24 hours, the standard was finally adopted).
  - Ttarget value (the return series for each asset) was normalized using StandardScaler.

THIRA

Methodology: Features selection

- Method: A Majority vote selection process based on mutual information

- Process:
  - A 5-fold time series cross-validation was used on the training set.
  - In each fold, the mutual information score between the features and the target was calculated for each asset, and the top 20 features were selected.
  - All cross-validation results were aggregated, and the 20 features with the highest frequency across all assets and time folds were selected as the final feature set.
  - Reasoning: This process ensures that the selected features are not only valid once, but also have universal predictive capabilities across different time periods and assets, greatly improving the stability of the feature set.

THIRA

Part 2:Comparative Analysis & Deep Dive

Landscape of Explored Models

- A wide range of architectures were systematically tested to find the best fit for our data:

- Tree-based models: LightGBM, XGBoost
- Neural network model: GRU, Stacked GRU, Bi-directional GRU (BiGRU), LSTM, CNN, Temporal Convolutional Network (TCN), MLP
- Attention-based sequence models: Encoder-only time series transformer, Temporal fusion transformer (TFT) , Informer (efficient long-sequence transformer)
- Due to insufficient RAM, the complete results could not be obtained and were therefore not included in the analysis.

Overall Performance

» Annualized Sharpe Ratio for each model across the different backtesting periods:

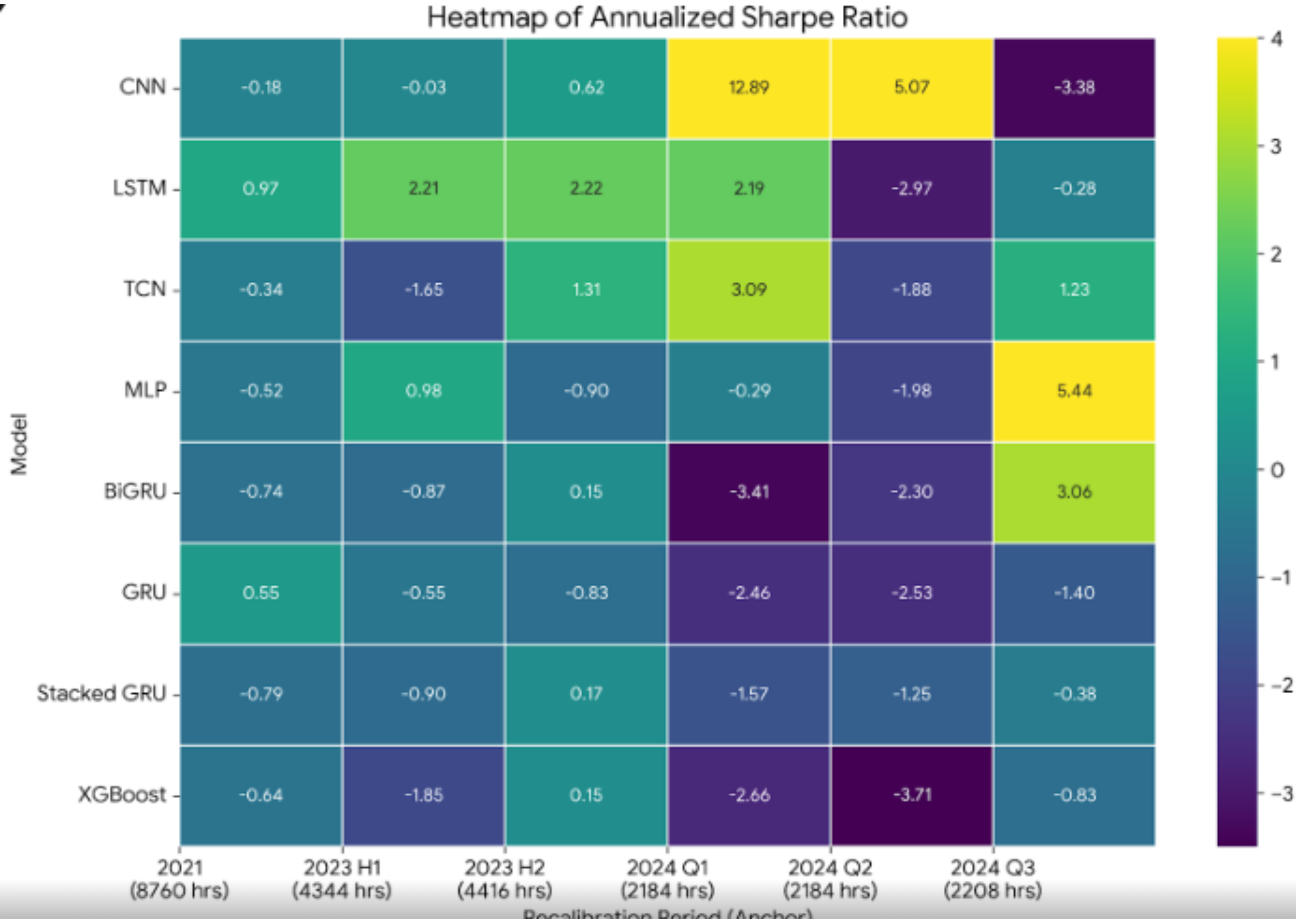
Model	2022	H1 2023	H2 2023	Q1 2024	Q2 2024	Q3 202
CNN	-0.27	-0.03	0.45	**3.89**	1.57	-3.12
LSTM	0.54	1.07	**1.76**	0.94	-2.29	-0.16
TCN	-0.25	-3.20	1.07	1.47	-1.22	0.54
MLP	-0.87	0.50	-0.75	-0.18	-1.22	**2.44**
BiGRU	-0.80	-1.20	0.13	-3.35	-1.11	1.46
GRU	0.67	-0.73	-0.97	-2.66	-2.10	-0.53
Stacked GRU	-1.19	-1.27	0.09	-1.06	-1.10	-0.26
XGBoost	-0.72	-1.79	0.08	-2.02	-2.28	-0.38

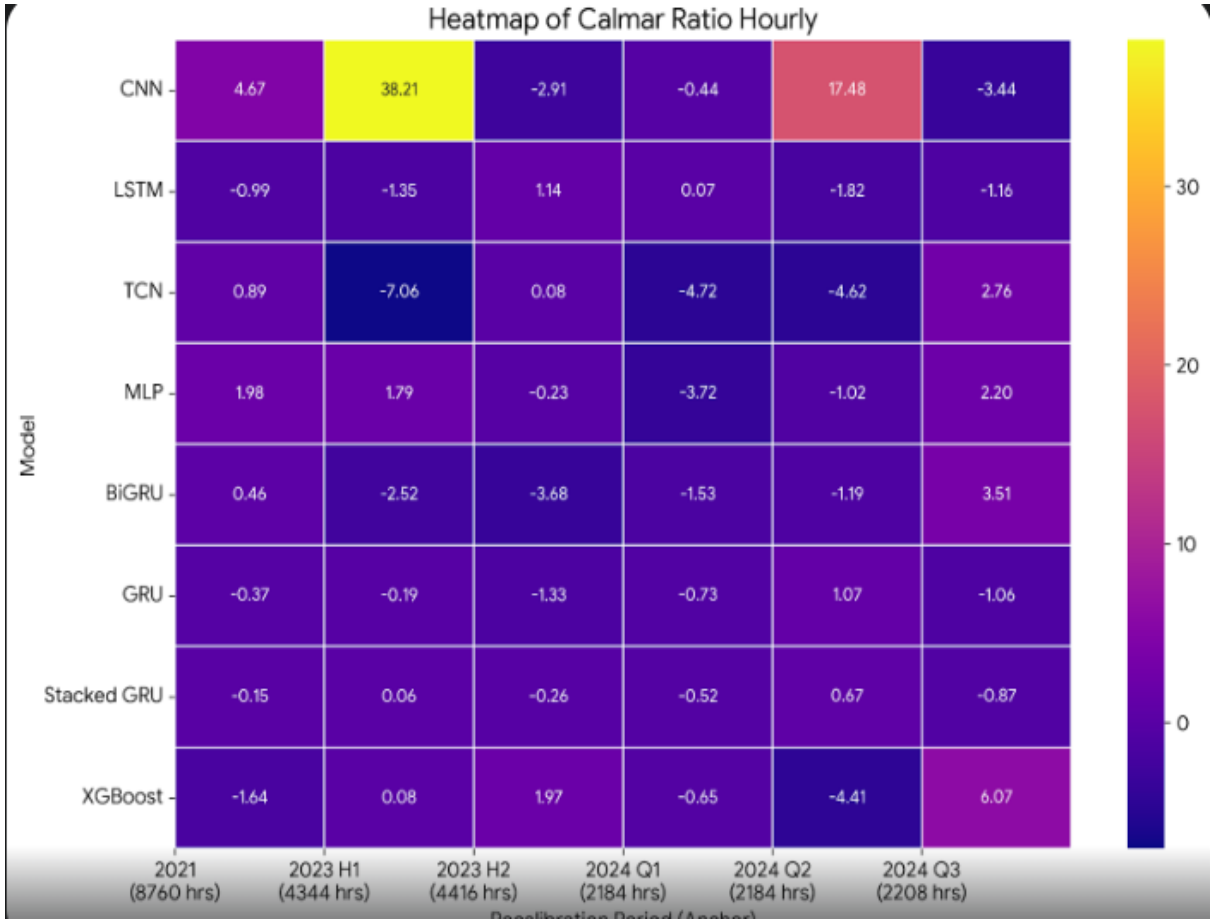
› Key Insight: While no model was universally profitable, deep learning models (CNN, LSTM) consistently found periods of high profitability.

› XGBoost never achieved this profitability behavior.

THIRA

Performance Landscape





THIRA

Qualitative Assessment

- CNN/ TCN : CNN-Highest Potential. Achieved the highest SR but also showed significant inconsistency. TCN did not show the explosive power of CNN, and its overall performance was more moderate.

- LSTM: Most Consistent Performer. Delivered positive Sharpe ratios for three consecutive periods. A more reliable candidate.

- GRU and its variants (GRU, Stacked GRU, BiGRU): Overall performance was poor. The BiGRU occasionally found success, but overall performance was still poor.

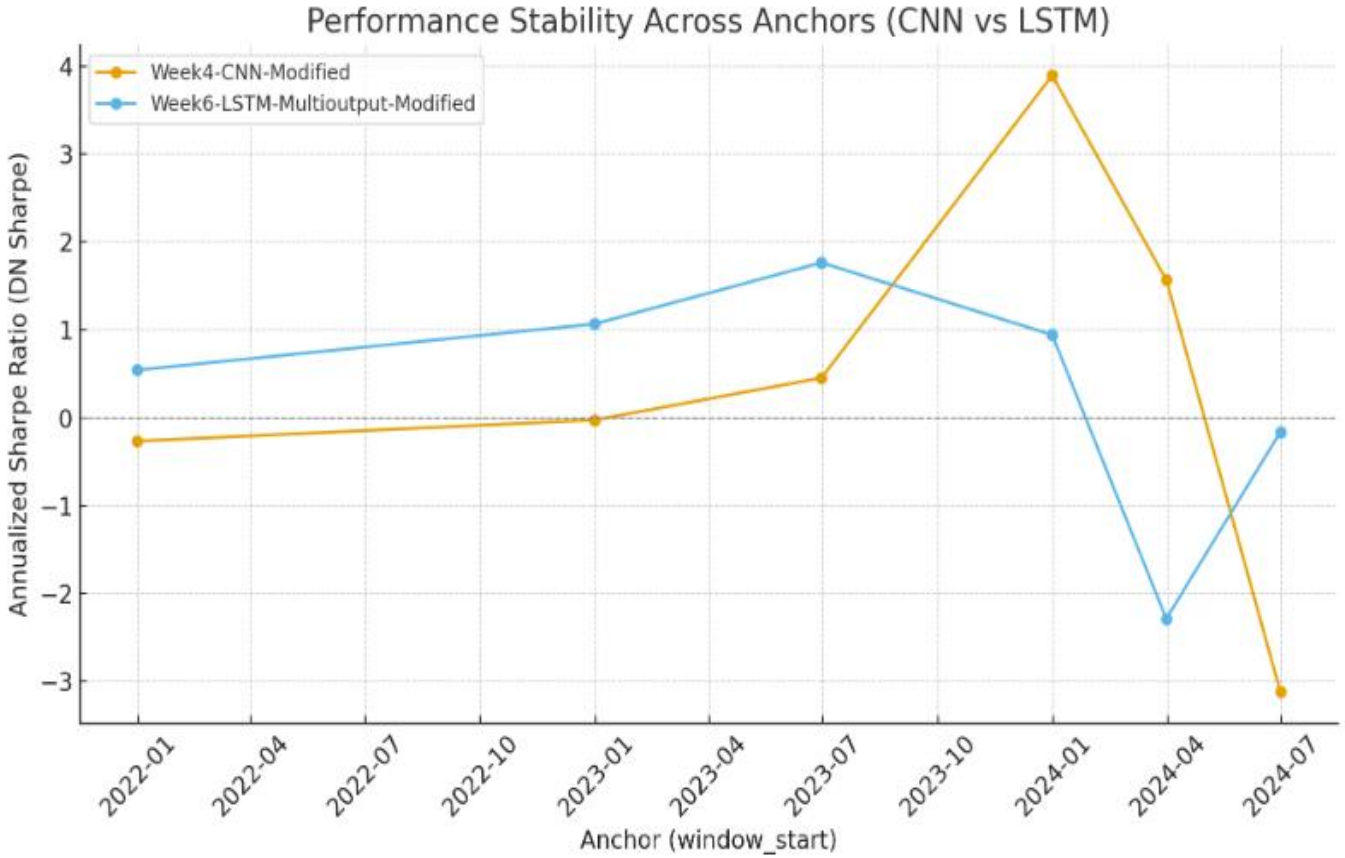
- Non-Sequential Models :

MLP: It achieved an amazing Sharpe ratio of 2.44 in the third quarter of 2024, but has underperformed at other times. XGBoost performs the worst. It seems to struggle with time series forecasting tasks.

THIRA

Head-to-Head: CNN vs. LSTM

- CNN: The Sprinter
  - Excels at capturing short-term, local patterns.
  - This likely explains its explosive result in the market of Q1 2024 but not stable.
- LSTM: The Marathon Runner
  - Better at understanding longer-term sequential dependencies. This led to its most consistent period of profitability.



Trade-off: Seems CNN can capture short-term opportunities, while LSTM can have more stable and consistent signal generation.
If we choose a single model, we need to make trade-offs.

THIRA

Deconstructing the CNN‘s Success and XGBoost’s struggle

- The CNN model's Sharpe Ratio of **3.89** with a max drawdown of only **-0.18%** in Q1 2024 is an exceptional result.
- My personal hypothesis
  - The market during this period likely exhibited strong, clear directional trends.
  - The mechanisms of CNN (local receptive field, weight sharing, high-frequency sensitivity) are just able to effectively extract these predictive patterns from the time series data of this period.

- XGBoost Consistent Underperformance**:** failed to produce a single profitable backtesting period.
- My personal hypothesis
  - Core takeaway: The XGBoost mechanism is designed for independent features and cannot understand temporal order. To make the data compatible, we had to "flatten" the 24-hour time series into 480 independent feature columns. This process destroys the crucial, time-ordered relationships between data points, which is the exact structural information that sequence models like CNNs and LSTMs are built to exploit.

THIRA

Key Challenges & Solutions

- Model performance decay in new market regimes is a persistent issue, as seen in the fluctuating Sharpe ratios for all models.
  - Try using hybrid models.
  - My initial idea is to develop and backtest a Convolutional-Long Short-Term Memory (Conv-LSTM) model to combine the advantages of CNN and LSTM, and then try to see if it performs stably in new market environments.

- Extremely high MAPE values across all models suggest poor accuracy in predicting exact return values.
  - Simplify the task to direction prediction (classification): This forces the model to focus on the "correct direction" that we actually care about, perfectly aligning its objective with our final trading decision.

Conclusion & Actionable Recommendations

» Conclusion: By backtesting and comparing 8 models, we confirmed that models that can understand time series structure (such as CNN and LSTM) have significant performance advantages for our tasks. The CNN architecture has the highest profit potential and should be explored and improved upon.

- Recommendation 1 (Short-Term):Deep-dive into the key drivers during the CNN's peak performance in Q1 2024.
- Recommendation 2 (Medium-Term): Explore methods to enhance models' robustness, enabling stable performance across a wider variety of market regimes.
- Recommendation 3 (Short-Term): Obtain more computing power and RAM to train the three attention mechanism frameworks I have built, and analyze the backtest results to determine whether the attention mechanism is suitable for our task and data.
- Recommendation 4 (Long-Term): Advanced Feature Engineering: Explore more data sources to create more powerful predictive features.

THIRA

Thank You & Q&A