A Systematic Evaluation of Machine Learning Architectures for Cryptocurrency Return Prediction

Ziwei Zhan

Internship Final Presentation

Thira Lab

THIRA

Part 1:Setting the Stage

Objective: Search for True Alpha

- **Primary Goal:** To identify machine learning models that generate predictive signals strong enough to be profitable in a **dollar-neutral portfolio**.

- **Why Dollar-Neutral?** This strategy aims to neutralize market exposure (beta) and isolate the model's true predictive edge (alpha).

- **Key Question:** Which model architecture best isolates alpha, delivering consistent risk-adjusted returns when market direction is removed from the equation?

THIRA

Evaluation Framework

- **Universe:** 10 diverse models tested (CNN, GRU, KNR, KRR (linear, rbf), LightGBM, LSTM, SVR (linear, rbf), TCN).
- **Data:** Top 3 liquid coins (BTC, ETH, BNB) on an hourly frequency.
- **Feature Selection:** Dynamic selection based on Spearman's Rank Correlation at each recalibration.
- **Backtesting:** Walk-forward recalibration at quarterly and semi-annual intervals.
- **Evaluation:** A dual-strategy backtest comparing "Raw" vs. "Dollar-Neutral" performance.

All models are portfolio-based models.

Feature Selection Method

- **1. Time-series cross-validation**
- Applied a 5-fold **TimeSeriesSplit** on the training set.
- Ensures feature relevance is evaluated consistently across different historical segments.
- **2. Information Coefficient (IC) calculation**
- For each fold, computed the **Spearman correlation (IC)** between every feature and the target return of the given coin.
- This measures **monotonic predictive power** of each feature.
- **3. Top-quartile filtering**
- Within each fold, kept only features whose **absolute IC exceeded the 75th percentile** (top 25% most predictive features).
- **4. Feature stability check**
- Collected selected features across all folds.
- Counted how often each feature was selected.
- Retained only features that appeared in **≥3 folds** (i.e., stable predictors).
- **5. Final feature set**
- Combined stable features across all coins → built a unified set of predictors for model training.

Selected features per recalibration date:

- 2021-12-31 → **75 features**
- 2022-12-31 → **77 features**
- 2023-06-30 → **77 features**
- 2023-12-31 → **71 features**
- 2024-03-31 → **80 features**
- 2024-06-30 → **70 features**
- 2024-09-30 → **78 features**

Average IC = 0.0081

THIRA

Hyperparameter Search Range

- **1. Convolutional Neural Network (CNN)**
- Filters: 32 – 128
- Kernel size: 3 – 7
- Dropout rate: 0.2 – 0.5
- Learning rate: 1e-4 – 1e-2 (log-uniform)

- **2. Gated Recurrent Unit (GRU)**
- Units: 32 – 256
- Dropout rate: 0.1 – 0.5
- Recurrent dropout rate: 0.1 – 0.5
- Learning rate: 1e-4 – 1e-2 (log-uniform)

- **3. k-Nearest Neighbors Regressor (KNR)**
- Number of neighbors: 2 – 30
- Weights: {uniform, distance}
- Metric: {minkowski}

- **4. Kernel Ridge Regression (KRR)**
- Alpha: 1e-4 – 10 (log-uniform)
- Gamma: 1e-4 – 10 (log-uniform)
- Kernel: linear, rbf

- **5. LightGBM**
- Number of estimators: 100 – 1200
- Learning rate: 0.01 – 0.3 (log-uniform)
- Max depth: 3 – 12
- Number of leaves: 15 – 255
- Min child samples: 5 – 50
- Subsample: 0.5 – 1.0
- Column sample by tree: 0.5 – 1.0

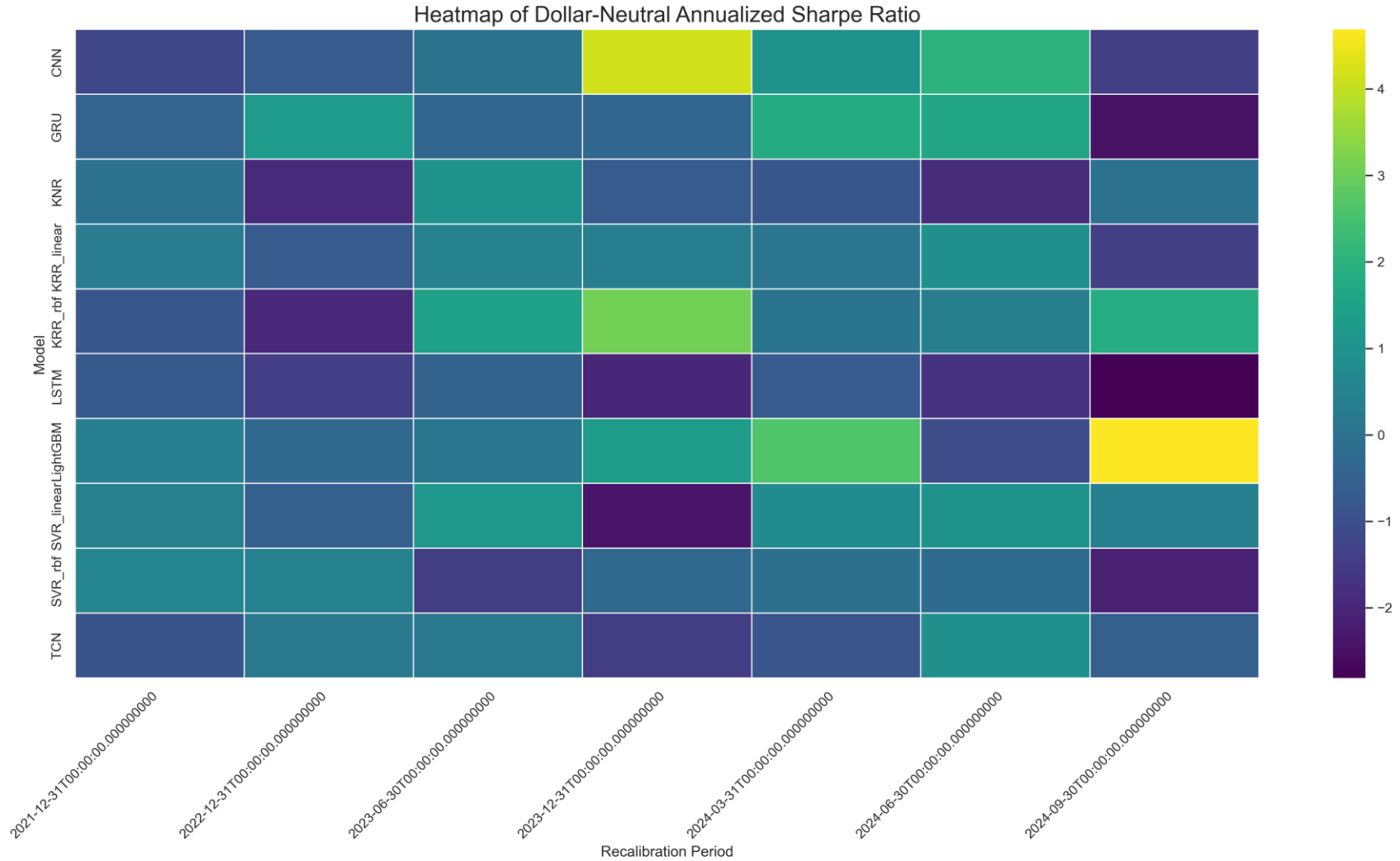
- **6. Long Short-Term Memory (LSTM)**
- Units: 32 – 256
- Dropout rate: 0.1 – 0.5
- Recurrent dropout rate: 0.1 – 0.5
- Learning rate: 1e-4 – 1e-2 (log-uniform)

- **Support Vector Regression (SVR)**
- C: 0.1 – 100 (log-uniform)
- Epsilon: 0.001 – 1.0 (log-uniform)
- Gamma: 1e-4 – 1.0 (log-uniform)
- Kernel: linear, rbf

- **8. Temporal Convolutional Network (TCN)**
- Number of filters: 32 – 128
- Kernel size: 3 – 8
- Dropout rate: 0.1 – 0.5
- Learning rate: 1e-4 – 1e-2 (log-uniform)
- Dilations: {(1,2,4), (1,2,4,8), (1,2,4,8,16)}

Part 2:Comparative Analysis & Deep Dive

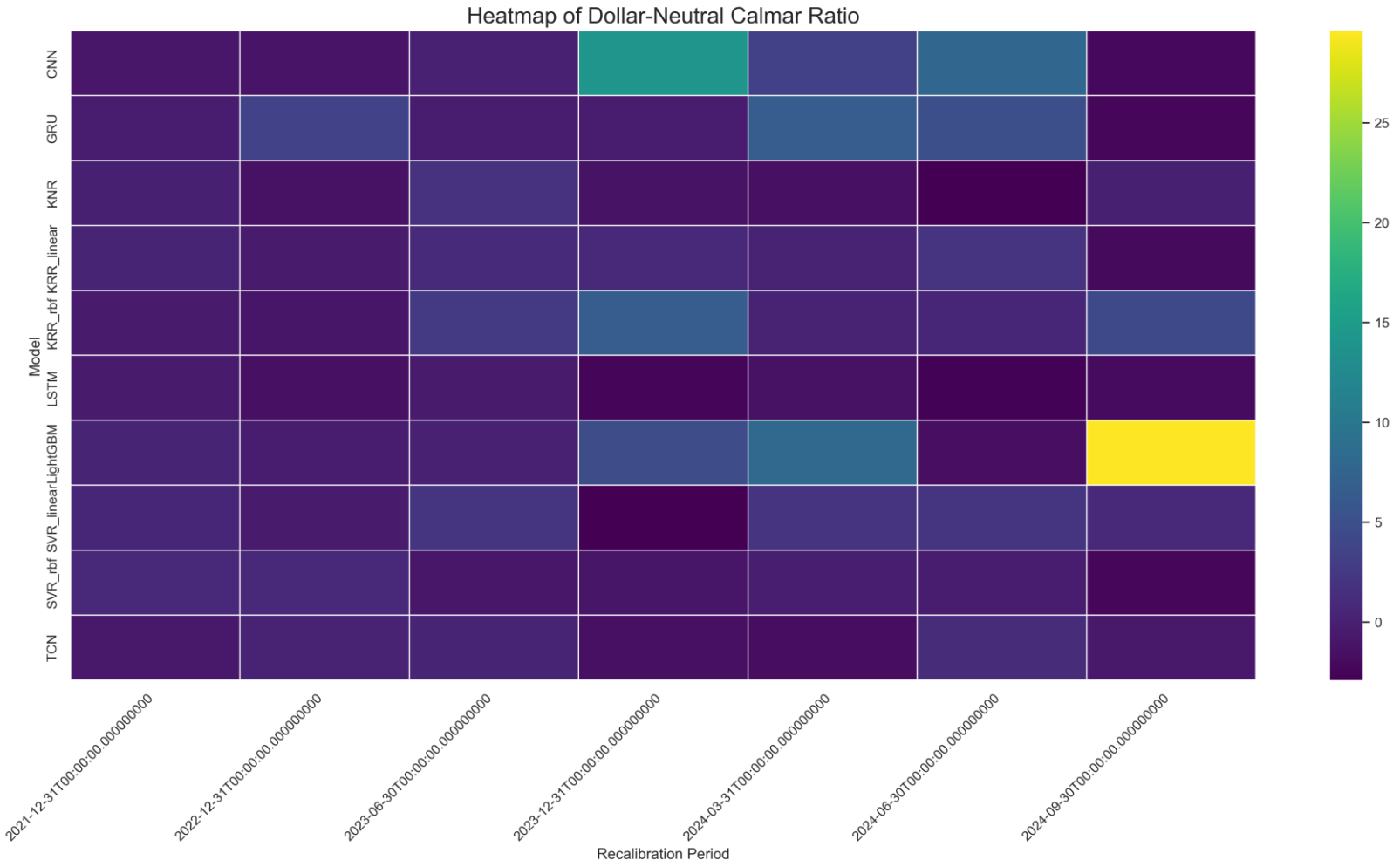
Performance Overview: Dollar-Neutral Annualized Sharpe Ratio



This provides an immediate visual summary of model performance, highlighting which models excelled (green) and which struggled (blue/purple) across different market regimes.

THIRA

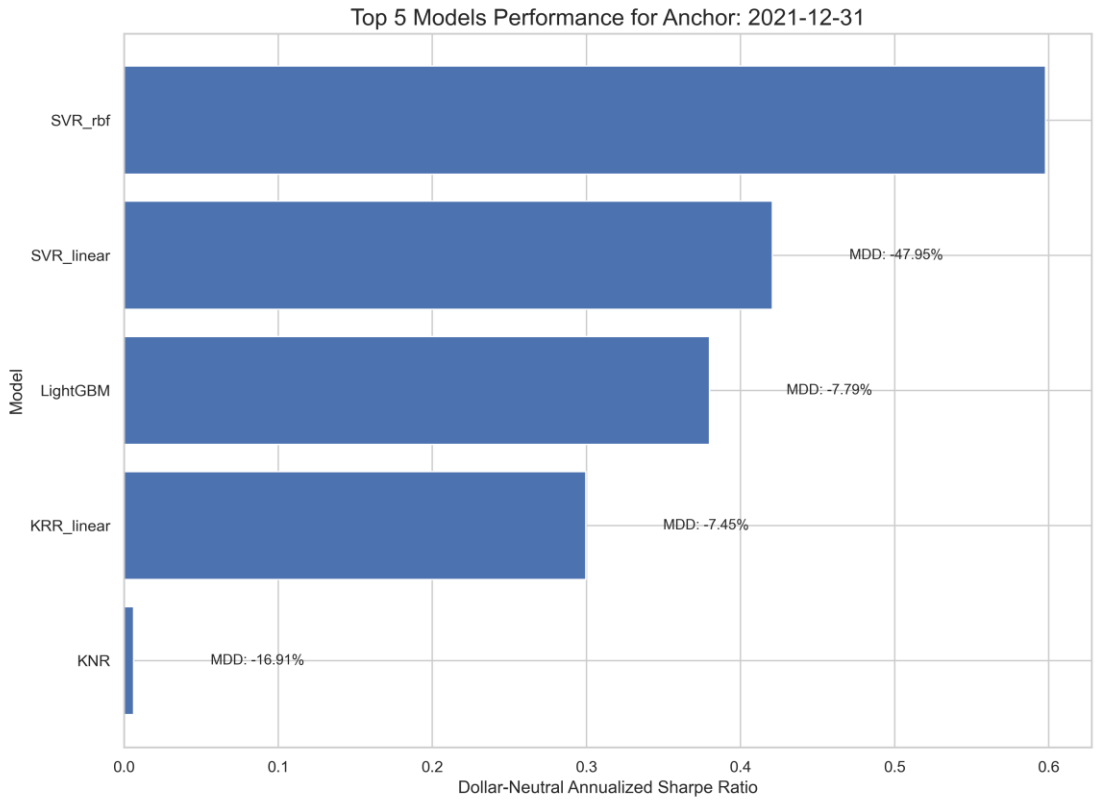
Risk Profile Overview: Dollar-Neutral Calmar Ratio

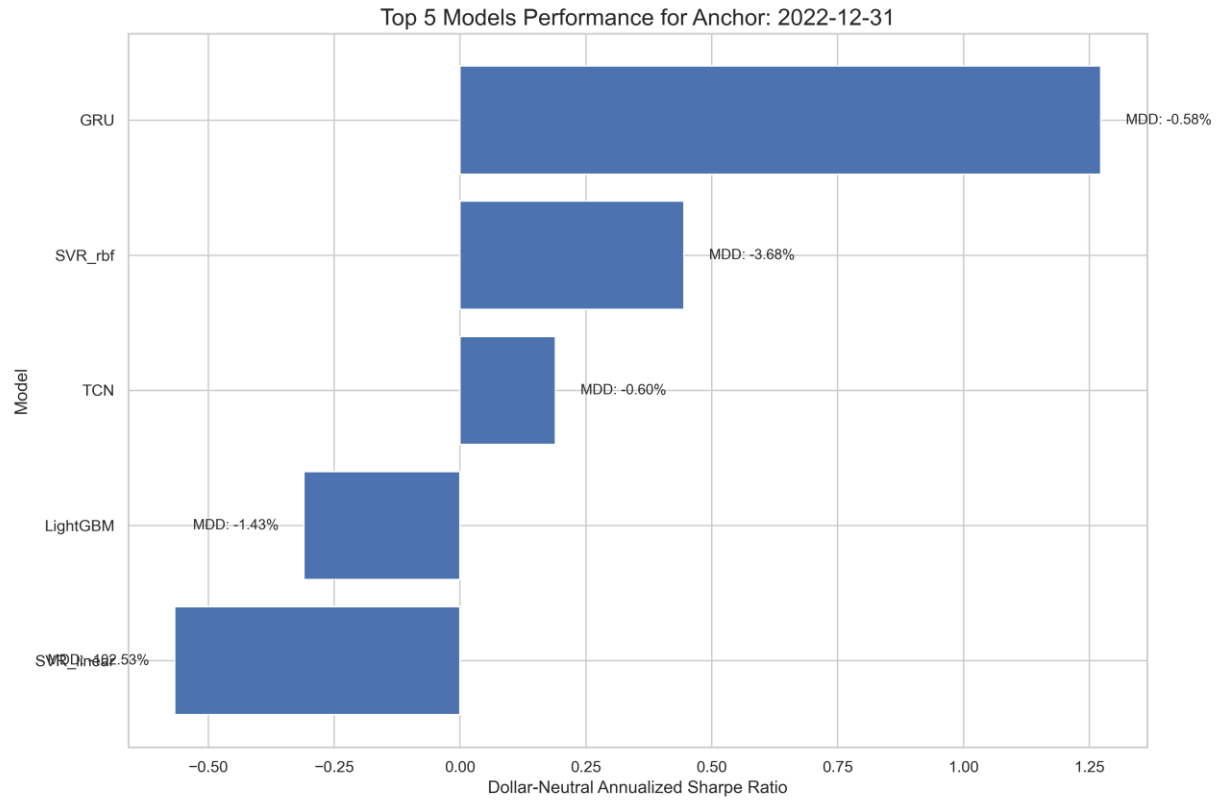


This view focuses on return relative to maximum drawdown, giving a clearer picture of risk-adjusted performance during each model's most challenging periods.

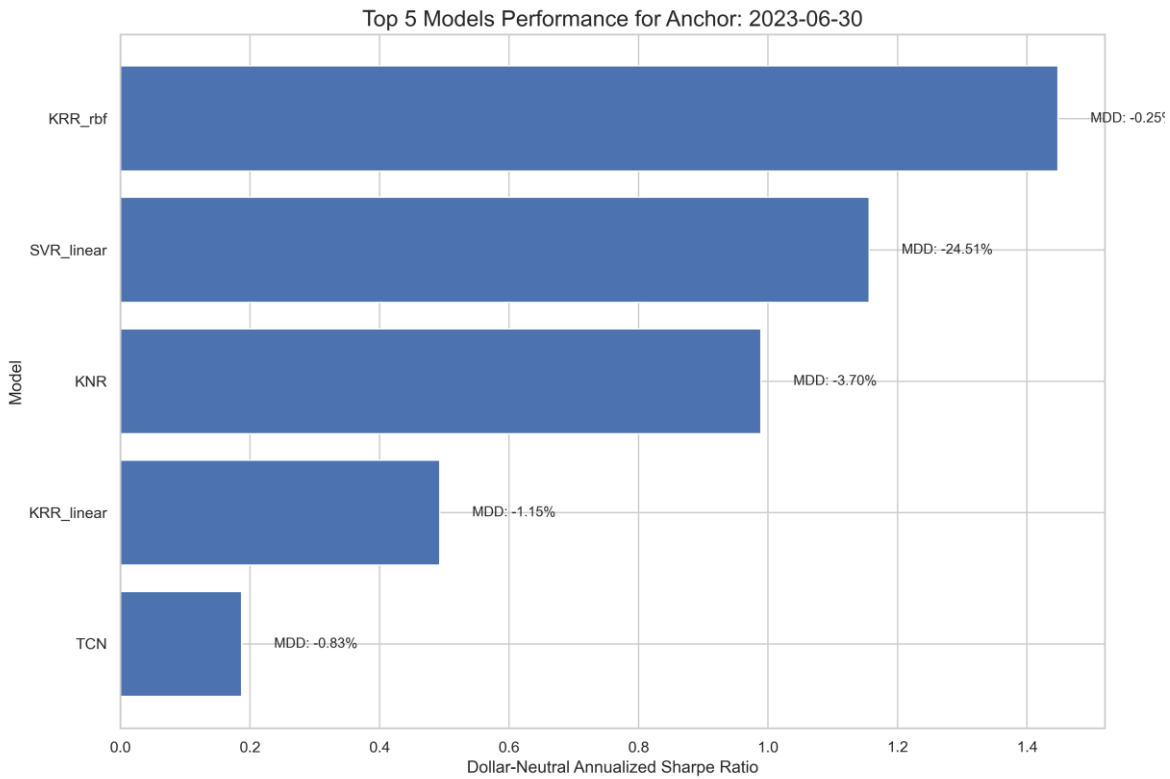
THIRA

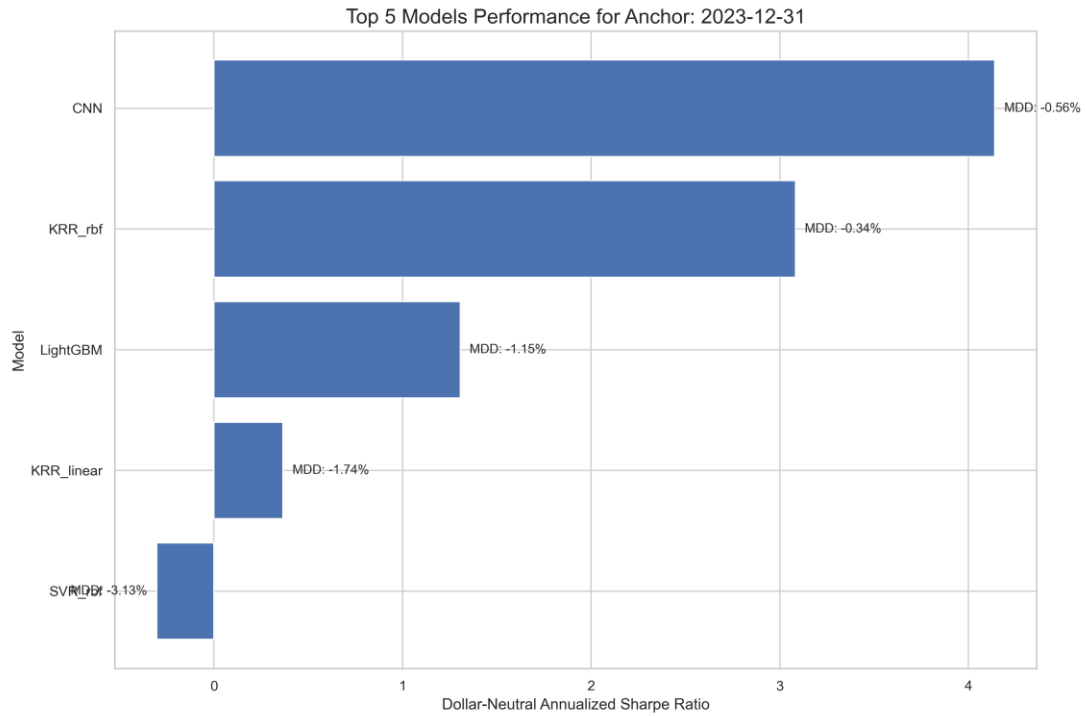
Per-Anchor Ranking





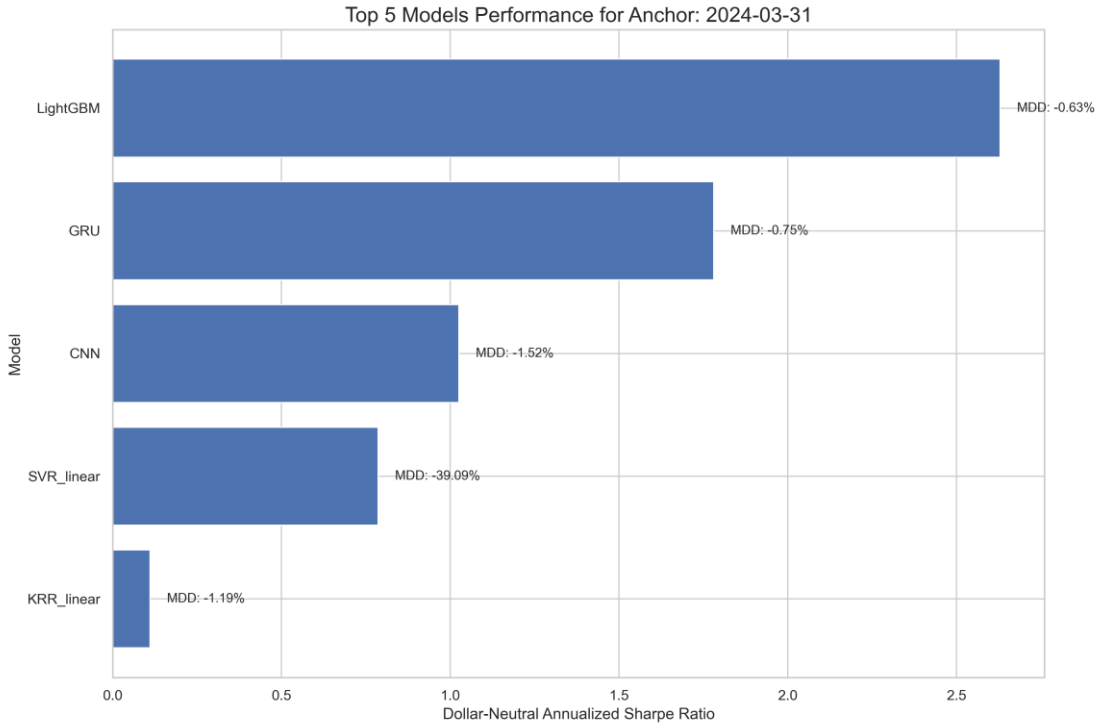
Per-Anchor Ranking

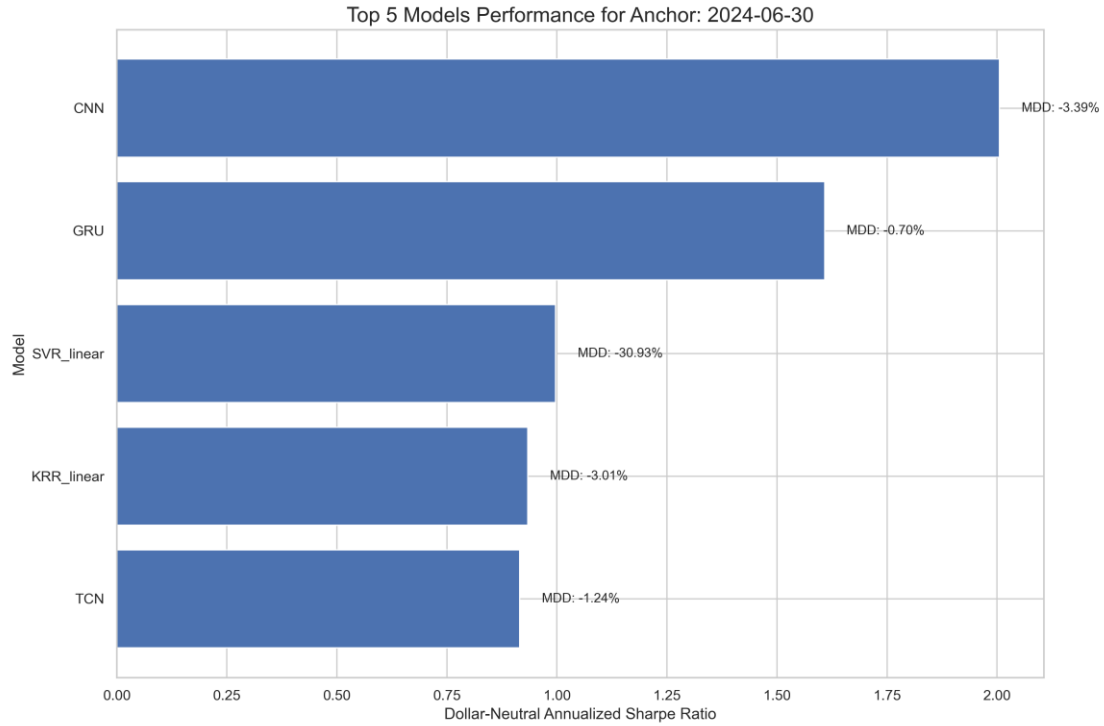




THIRA

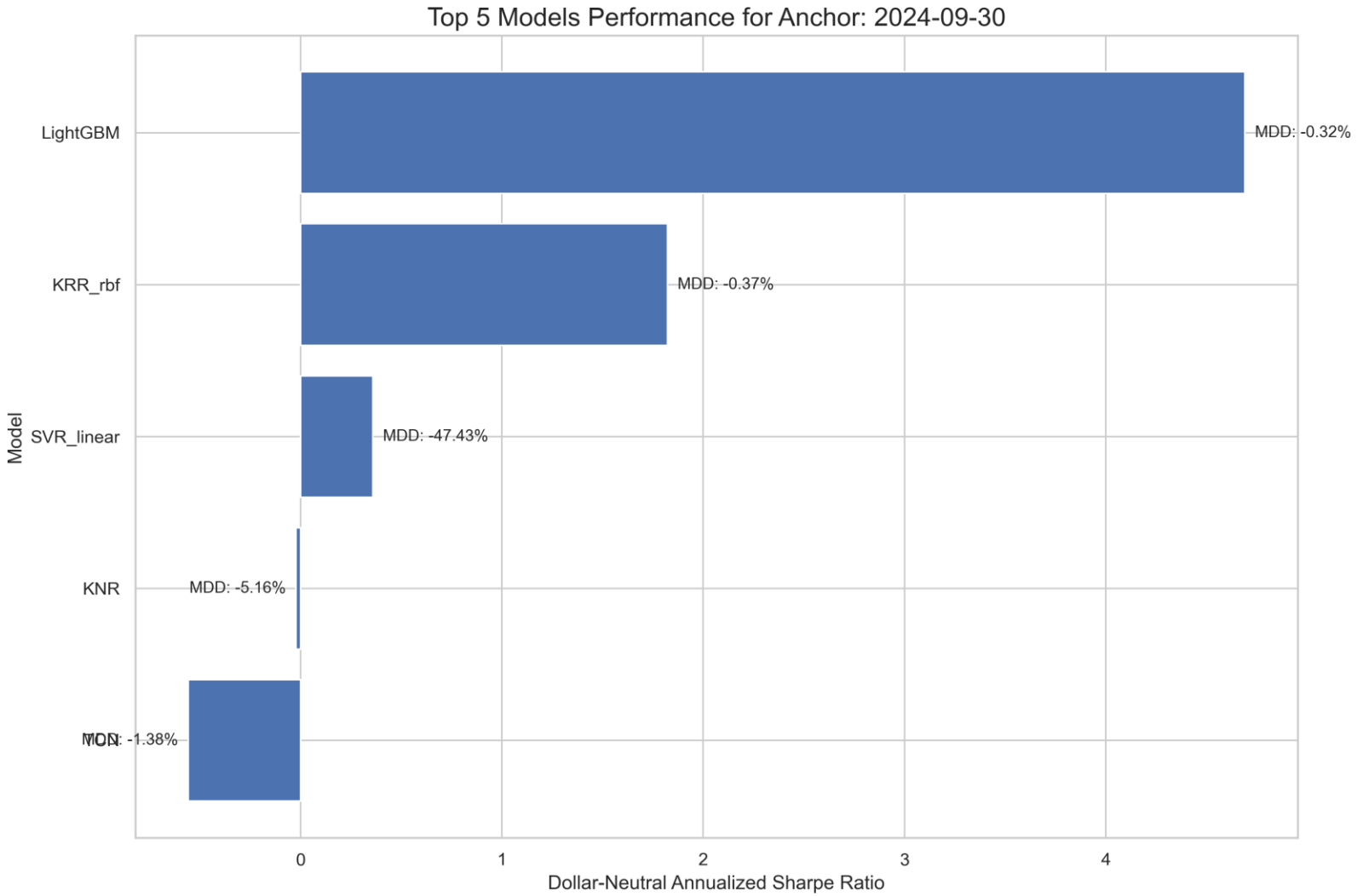
Per-Anchor Ranking





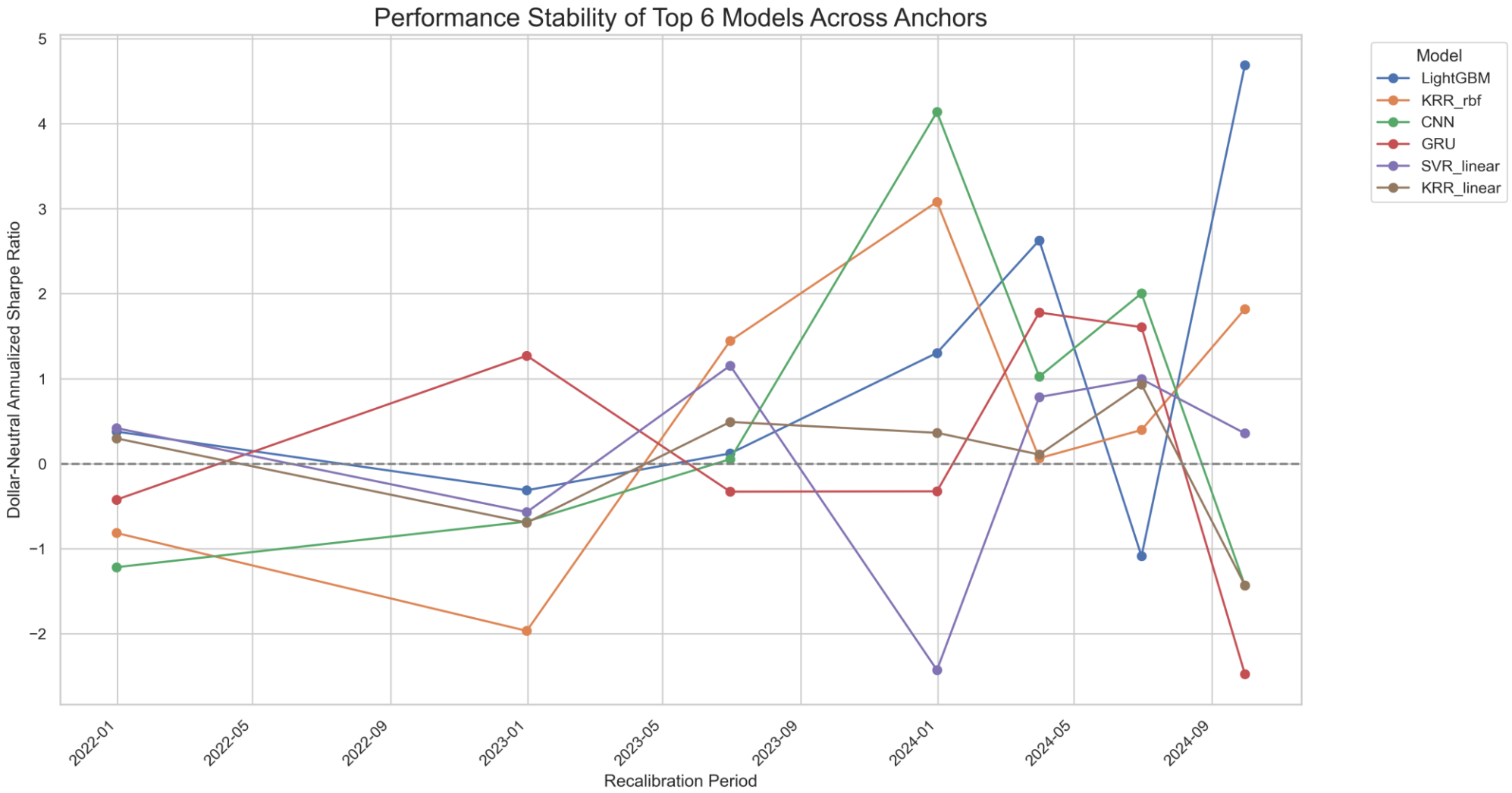
THIRA

Per-Anchor Ranking



THIRA

Models Consistency



THIRA

Models Consistency


```
--- Model Rank Stability ---
(Lower mean and std are better)
             mean   std
Model
LightGBM     3.71  2.56
SVR_linear   4.14  2.79
GRU          4.86  3.18
KRR_linear   4.86  1.21
CNN          5.00  3.42
KRR_rbf      5.00  3.42
SVR_rbf      5.71  3.25
TCN          6.43  2.57
KNR          6.71  2.75
LSTM         8.57  0.98
-------------------------
```

- **Model Rank Stability (based on Annualized Sharpe Ratios)**
- Models are ranked within each recalibration period by their Annualized Sharpe Ratio.
- The table reports **average rank (mean)** and **rank variability (std)** across periods.
  - **Lower mean** → the model tends to rank higher on average.
  - **Lower std** → the model's relative position is more stable over time.
- Interpretation: A model with both low mean and low std is a consistently strong performer.

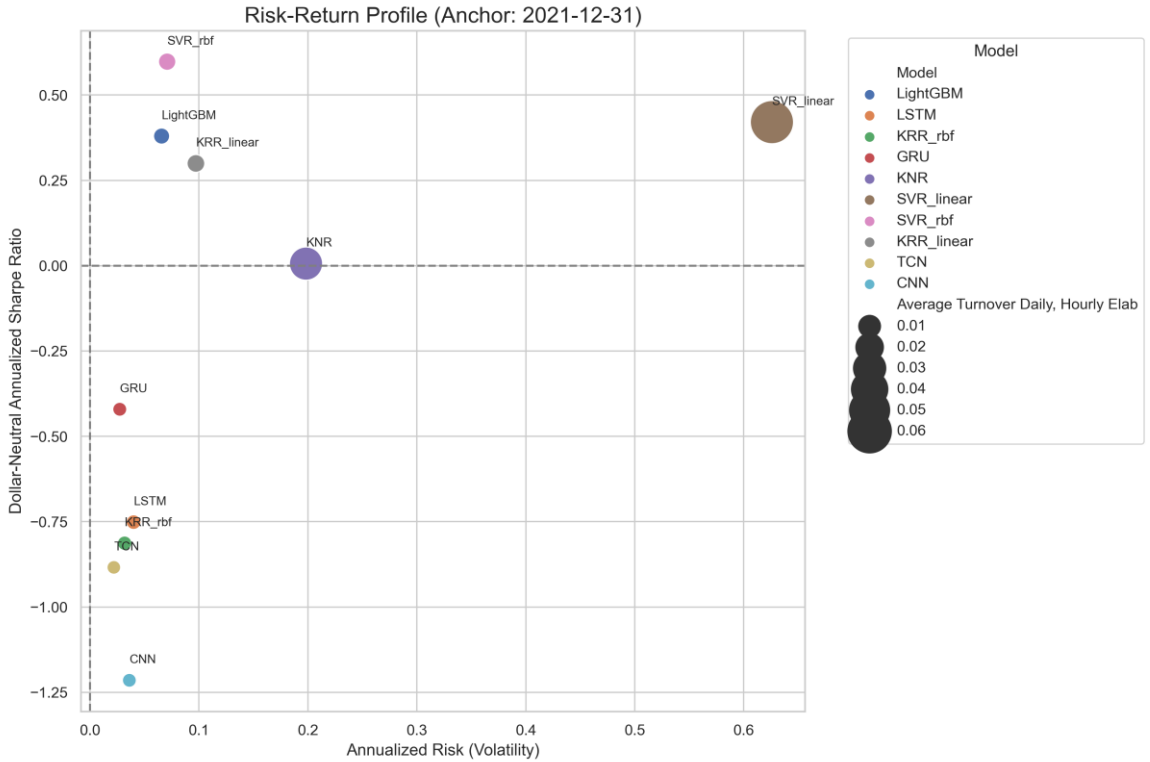
THIRA

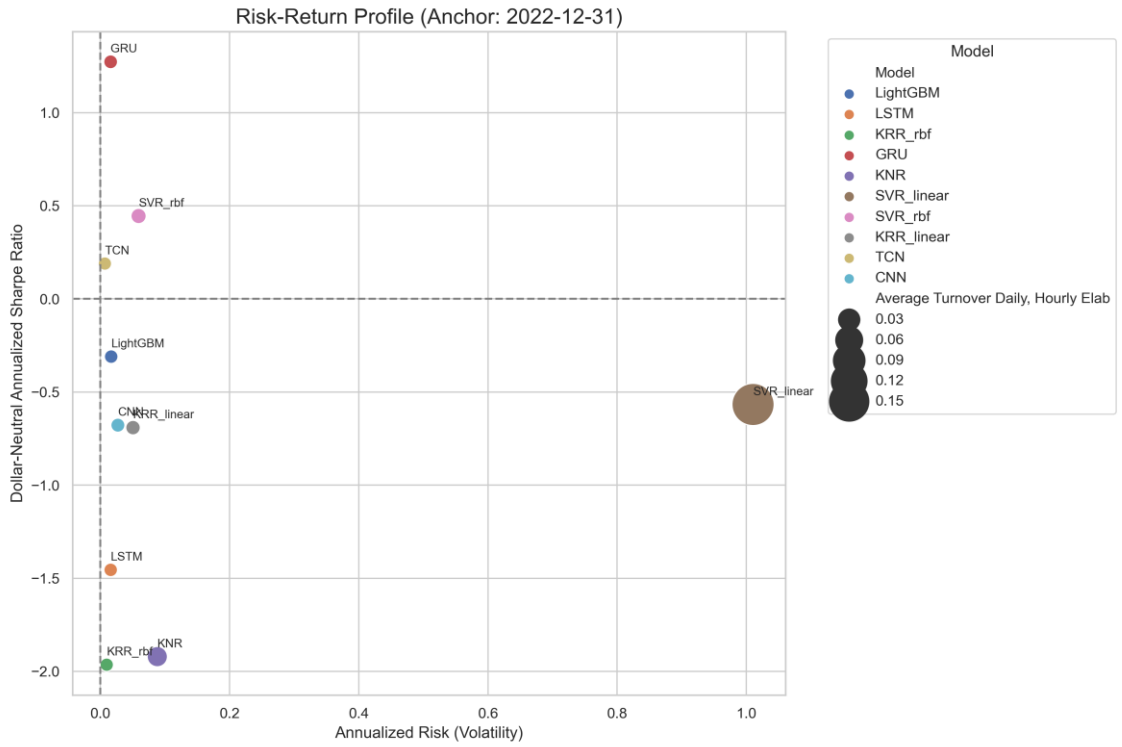
The Impact of Neutralization (ΔSharpe = DN Sharpe - Raw Sharpe)

It isolates true alpha generators (positive bars) from models whose performance was dependent on market direction (negative bars).

THIRA

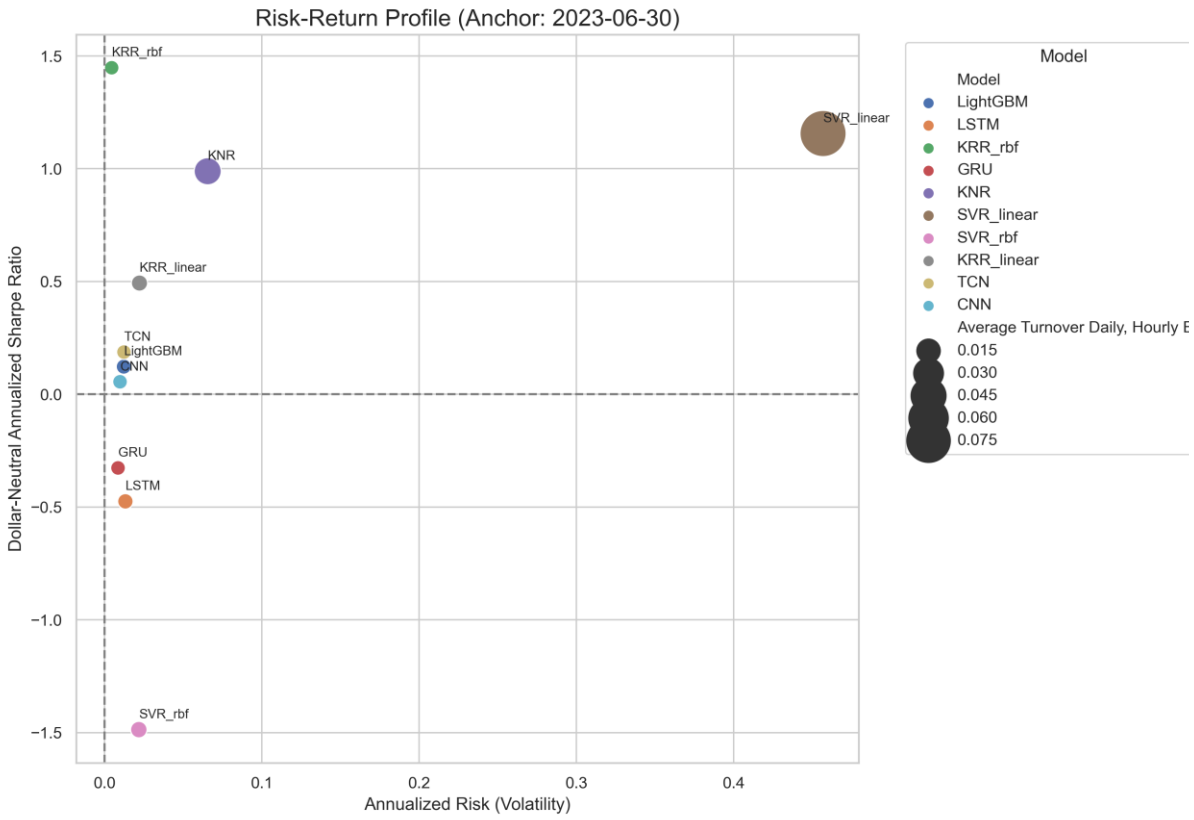
Risk-Return Profile for Top Models

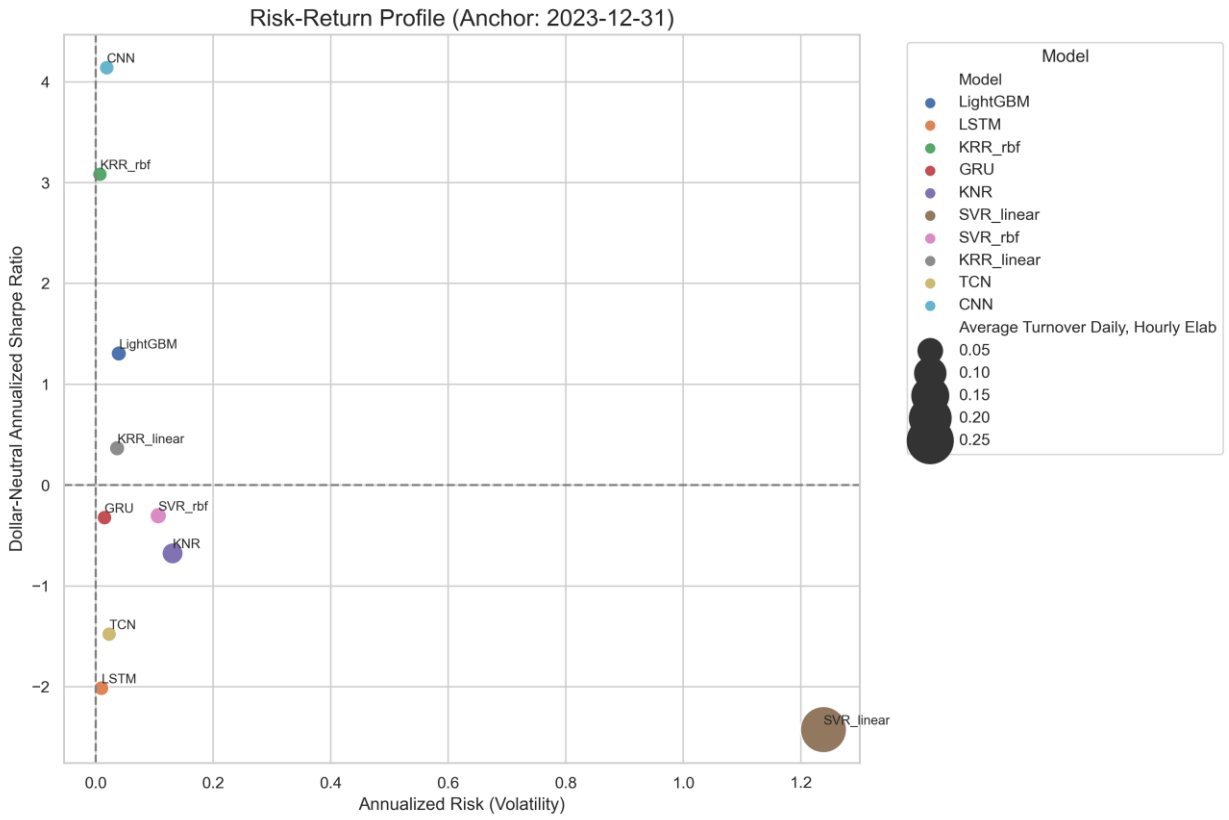




THIRA

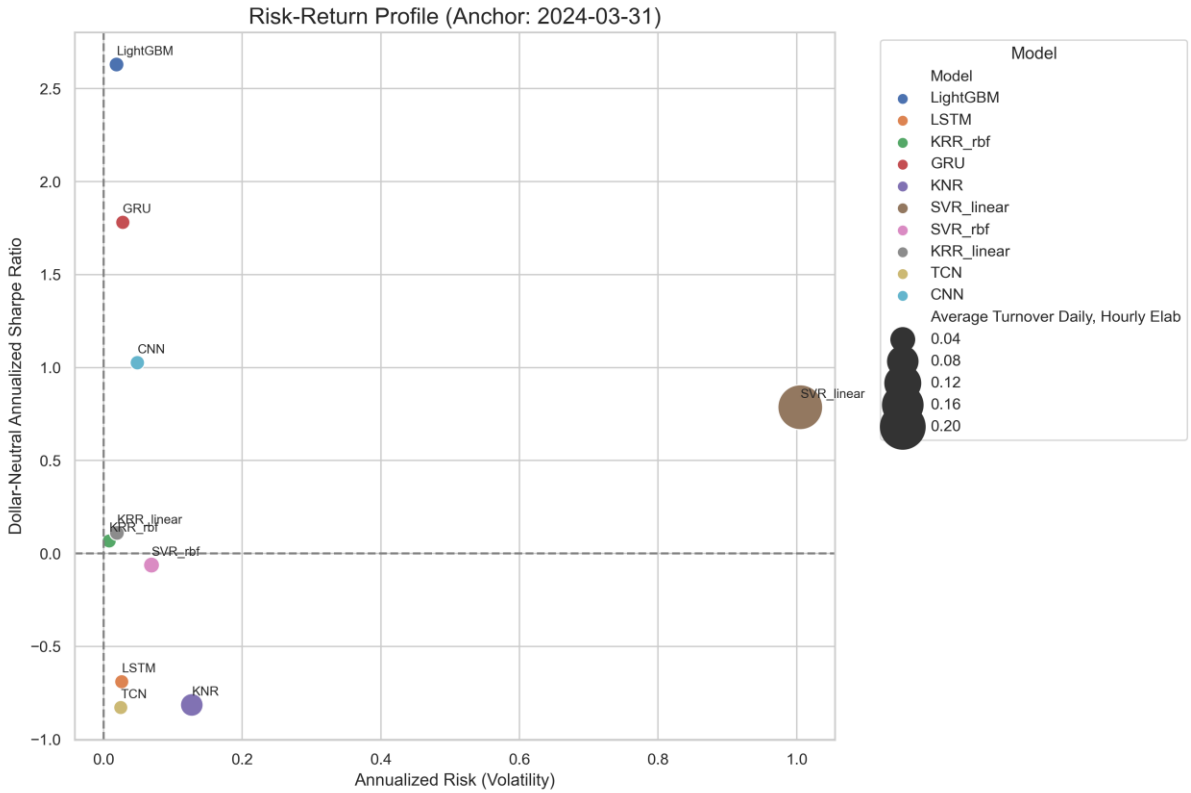
Risk-Return Profile for Top Models

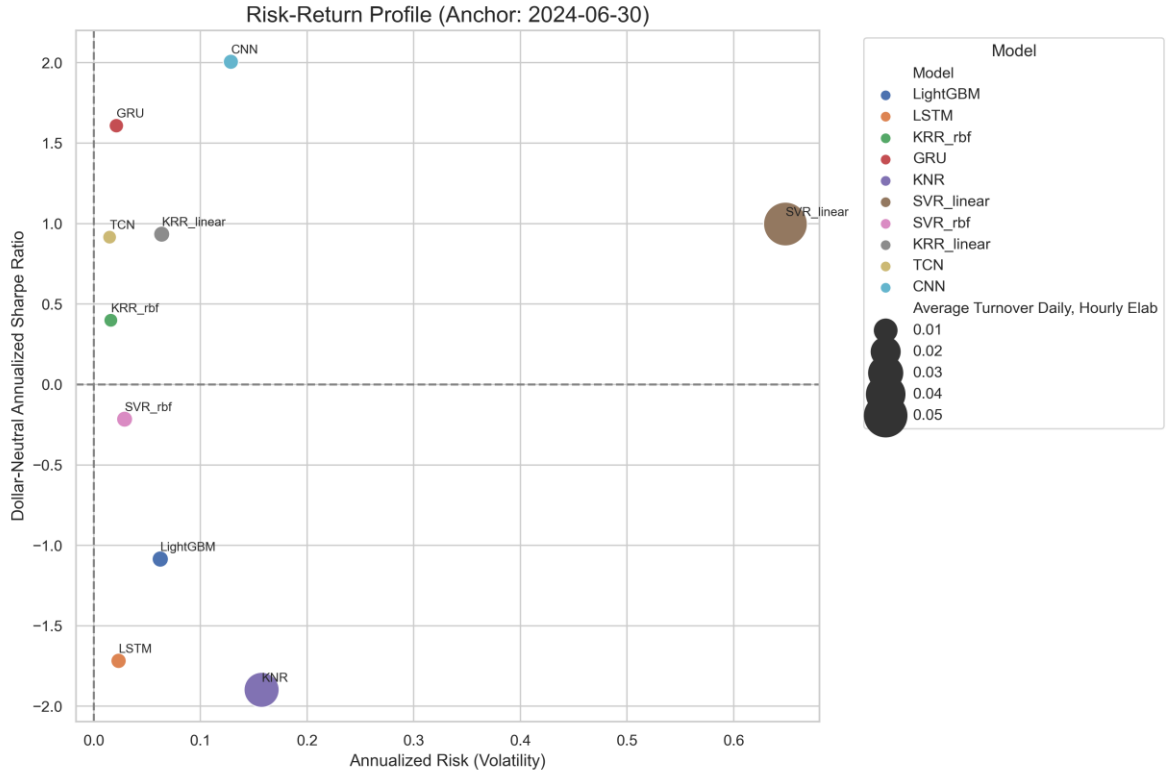




THIRA

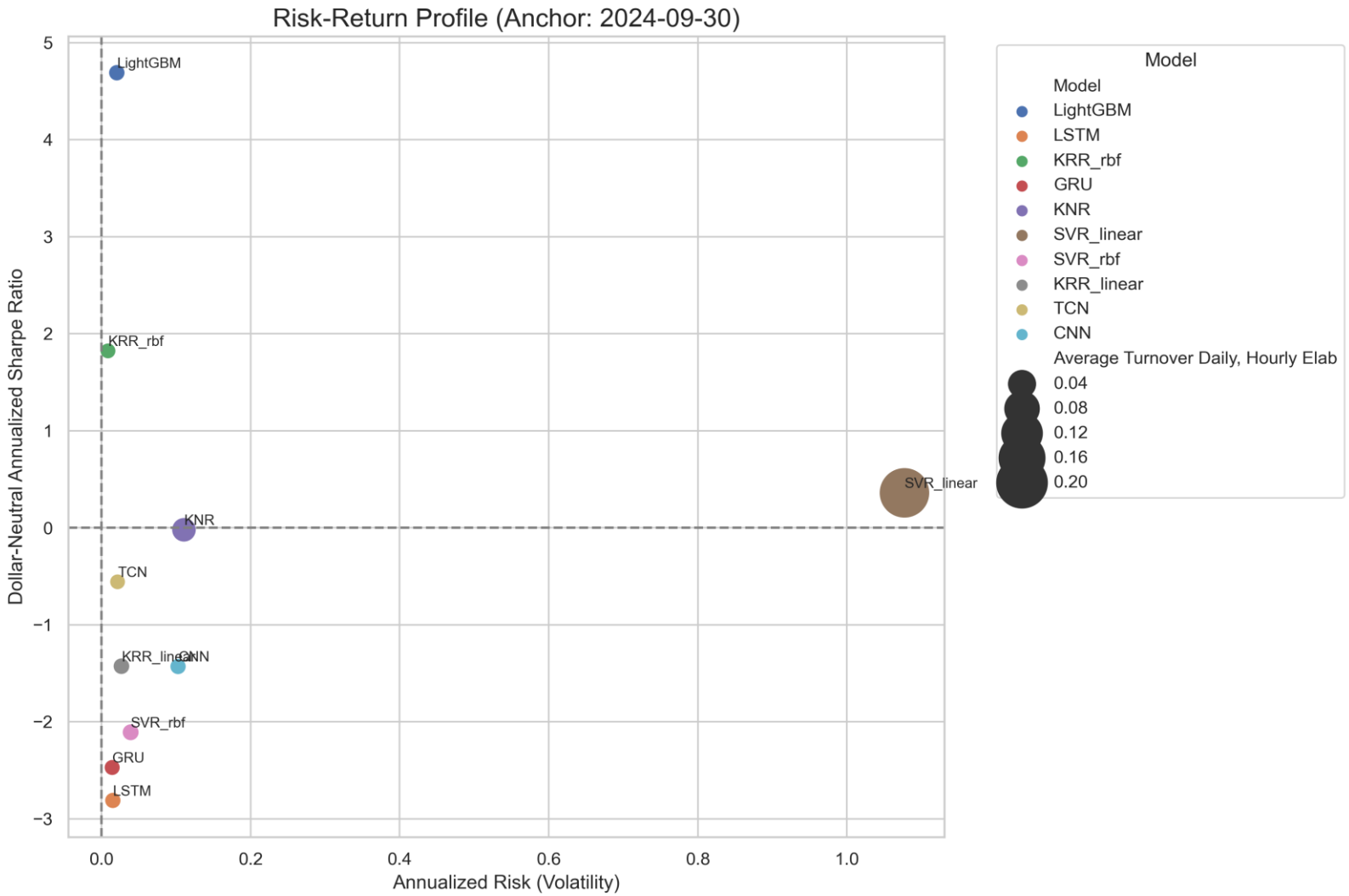
Risk-Return Profile for Top Models





THIRA

Risk-Return Profile for Top Models



Conclusion & Actionable Recommendations

- **Conclusions:**

- The systematic backtest identifies **LightGBM** and **KRR_rbf** as the top-performing models, consistently delivering higher risk-adjusted returns in a dollar-neutral strategy.
- While not always achieving the highest peak performance, **GRU or CNN** proved to be relatively *consistent* model, demonstrating the lower average rank and rank volatility across all backtesting periods.
- The dollar-neutral framework was highly effective at isolating true alpha. In the most recent period, **LightGBM** showed improved performance after neutralization, confirming its signals are valuable independent of market direction. In contrast, **KRR_rbf** lost Sharpe after neutralization, suggesting part of its raw edge relied on market drift rather than pure alpha.
- Conversely, complex sequence-based models like **LSTM** consistently underperformed in the dollar-neutral framework, indicating they were not suitable for this specific prediction task.

Conclusion & Actionable Recommendations

- **Recommendations:**

- **Short-Term: Prioritize & Blend**
- Focus immediate development on the **LightGBM** and **KRR_rbf** models. Explore blending the signals from these two top-performers to create a more robust meta-strategy.
- **Medium-Term: Utilize Consistency**
- Investigate the stable signals from the **GRU or CNN** model. It could serve as a reliable baseline or be blended with higher-performing models to reduce overall portfolio volatility.
- **Actionable Decision: De-Prioritize**
- De-prioritize further research on the **LSTM** architecture for this task due to its consistent and significant underperformance.
- **Future Research: Focus on Features**
- Focus future feature engineering on creating rich, interactive features. The success of tree-based (LightGBM) and kernel-based (KRR) models suggests that the relationships *between* features are more critical than their long-term sequence for this prediction horizon.

Future Directions

- **Target Redefinition**
- Current setup: **Regression on log-returns** (predict continuous future returns).
- Future direction: **Formulate as a classification problem**:
  - **+1** → Positive return
  - **0** → Neutral / small movement
  - **-1** → Negative return

- **Benefits**
- Aligns model output with **trading decisions** (long / flat / short).
- Potentially improves **signal robustness** in noisy financial data.
- Facilitates use of classification-focused metrics (Accuracy, F1, MCC) for evaluation.

Thank You & Q&A