Predictive Modeling and Optimization for Tax-Aware ETF/Mutual Fund Portfolio Re-balancing

Dhruv Arcot

dhruvag8@stanford.edu)

Introduction

- Mutual funds and ETFs are core investment instruments for modern investors helping with diversification and risk management.
- Portfolio re-balancing is a critical process for mutual funds operations, yet it is typically executed by fund managers using heuristic and subjective methods that are not tailored to individual investors nor are they completely data-driven.
- This project aims to build a tax-aware re-balancing system designed to outperform traditional mutual funds by maximizing net realized returns after accounting for capital-gains taxes.

Data & Features

There were multiple components to the project and the data and feature-engineering required for each is as follows

1. Stock Price Dataset
   - i) Data
     - Daily U.S. equity OHLCV data (2015–2025).
     - Includes all stocks appearing in selected sector mutual funds.
   - ii) Engineered Features
     - Rolling averages, volatility, lagged returns.
     - Momentum indicators (ROC, MACD), relative price ratios.
     - No look-ahead bias.
2. Mutual Fund Constituents
   - i) Data
     - Monthly holdings with tickers and portfolio weights.
     - Sourced from sector fund webpages.
   - ii) Sector Coverage
     - Each mutual fund was categorized based on their sector (Eg:- Technology, Financial Services, Pharmaceuticals and Medicine, Industrial e.t.c)

For the stock-prediction model, the dataset was divided into the following temporal segments:

- **Training:** 2015–2022
- **Testing:** 2022–2025

Models for Long-Horizon price predictions

The goal of the model was to perform long-horizon stock price prediction for a price H trading days in the future:

$$y_t = P_{t+H}, \qquad H \in \{180, 365\}$$

The prediction task was formulated as a regression problem, and the final evaluation metrics for model comparison were MSE and RMSE.

Several categories of models—Linear, Tree-based (XGBoost), and Sequential (RNN and LSTM) architectures—were evaluated. After conducting model diagnostics, Sequential Models, particularly LSTMs, were identified as the most effective candidates. The prediction problem for sequential models was defined as:

$$\hat{y}_t = f_\theta(x_{t-k:t}),$$

where f_θ is a sequence model operating on the previous 'k' observations with k treated as a tunable hyperparameter.

The objective was to train an individual learner for each stock to enable fine-tuned predictions rather than generalizing across multiple assets.

Training used the Adam optimizer (learning rate 1×10^{-3}), batch size 64, and mean squared error loss. Hyperparameters were selected using a focused grid search on the particular stocks validation-set.

Tax aware Optimization and Re-balancing strategies

Tax-Penalizing Strategy

Taxation effects were modeled at a lot level. For each asset i and its lot j, the following quantities are defined:

- $q_{i,j}$: shares held in lot j
- $c_{i,j}$: cost basis per share
- p_i: current market price
- $\Delta q_{i,j} \geq 0$: quantity sold from lot j
- $\tau_{i,j}$: applicable tax rate (short-term or long-term)

Realized gains from selling $\Delta q_{i,j}$ shares are computed as:

$$G_{i,j} = \Delta q_{i,j}(p_i - c_{i,j})$$

The tax owed on each lot is:

$$T_{i,j} = \tau_{i,j} \max(0, G_{i,j})$$

Total tax penalty across all assets and lots is:

$$T_{\text{total}} = \sum_i \sum_j \tau_{i,j} \max\left(0, \Delta q_{i,j}(p_i - c_{i,j})\right)$$

Re-balancing Optimization Strategy

Let w_i^{old} denote the current weight of asset i, w_i^{new} the proposed re-balanced weight, and $\hat{r}_i$ the predicted H-month return from the long-horizon model predictions.

The projected after-tax portfolio value under re-balancing is:

$$V_{\text{re-balance}} = V\left(1 + \sum_i w_i^{\text{new}} \hat{r}_i\right) - T_{\text{total}}$$

The projected value without re-balancing is:

$$V_{\text{no}} = V\left(1 + \sum_i w_i^{\text{old}} \hat{r}_i\right)$$

A re-balancing action is taken only if:

$$V_{\text{re-balance}} > V_{\text{no}}$$

Experiments

The experiments evaluated multiple model classes for long-horizon stock price prediction, focusing on understanding how performance changes as the forecast window increases. Using a grid search the ideal set of hyper-parameters were identified. Custom tuning of the sequence length alone was done in order to find the most optimal sequence length for all stocks.

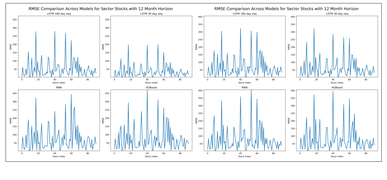

Figure 1. RMSE comparison of various models for 6-month and 12-month horizon windows

Model	Horizon (Months)	Avg RMSE
XGBoost	6	72.61
RNN	6	65.23
LSTM 180-Day Seq	6	52.23
LSTM 30-Day Seq	6	40.18
XGBoost	12	78.38
RNN	12	76.23
LSTM 180-Day Seq	12	66.25
LSTM 30-Day Seq	12	61.12

Table 1:Avg RMSE for each stock by models

In order to evaluate both the Long-Horizon Price Predictions and the Tax-Aware Optimization and Re-balancing Strategies, the following simulation framework was designed:

- Compute the baseline after-tax returns for each fund by simulating its monthly re-balancing schedule.
- Using the initial holdings and constituents, predict the next H-month returns using the selected long-horizon model.
- Apply the tax-penalizing strategy to evaluate whether a re-balancing action is beneficial.
- Compute realized capital gains and the resulting after-tax portfolio value.

In this manner, the framework directly compares the net returns of each mutual fund under its natural re-balancing behavior performed by the fund-manager against the proposed tax-aware re-balancing system.

Results & Discussions

Based on the RMSE charts and table it is observed that the LSTM models with a 30-day sequence window performed the best (Avg RMSE) across both the 6-month and 12-month horizon window. The model also performs better for the 6-month horizon as compared to the 12-month horizon.

This simulation framework was then executed on ten funds—five from the Technology sector and five from the Financial Services sector—selected based on their 2025 performance. The comparison on profits, before and after was computed and the results are as follows:

ID	Returns Before Tax	Returns After Tax
F001	-1.85%	0.45%
F002	-0.95%	0.91%
F003	1.11%	2.22%
F004	-2.10%	-0.25%
F005	-0.78%	1.15%

Table 2 Net returns improvement from tax-aware re-balancing (Technology Sector)

ID	Returns Before Tax	Returns After Tax
F006	-1.15%	0.98%
F007	-0.38%	1.25%
F008	-2.25%	-1.55%
F009	1.02%	2.48%
F010	-0.82%	1.35%

Table 3 Net returns improvement from tax-aware re-balancing (Financial Sector)

For 8 out of the 10 merchants positive improvement on net profits is observed and for 2 of them, positive returns are observed on the pre-tax returns indicating the long-horizon predictions and tax-strategy penalization is able to outperform the baseline fund. Overall, the tax-aware approach delivered a net uplift of **1.21**% across all evaluated funds, demonstrating its effectiveness for long-term, taxation-aware portfolio management.

Future

- Improve long-horizon prediction models.
- Feature enhancement by including market context into the models prediction
- Introduce an exploration–exploitation mechanism to identify new investment opportunities outside current fund holdings.
- Personalized re-balancing strategies based on tax slabs, risk appetite, and investment period.

References

- Moehle, N., Kochenderfer, M. J., Boyd, S., & Ang, A. (2021). *Tax-Aware Portfolio Construction via Convex Optimization*. arXiv:2008.04985.
- Garcia, D., Garan, N., & Singh, A. (2020). *Multi-Period Portfolio Optimization with Transaction Costs and Constraints*. Journal of Financial Optimization.
- Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation.