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

Stanford CS229 Project

Dhruv Arcot
Department of Computer Science
Stanford University
dhruva98@stanford.edu

1 Introduction

Mutual Funds are a core instrument in modern investing, offering diversified exposure across sectors to maximize returns while managing risk. However, the common practice of portfolio re-balancing, typically performed by financial advisors, remains largely heuristic, subject to human bias, and often overlooks the long-term tax implications of frequent trades. Realized capital gains from repeated re-balancing can substantially erode cumulative performance, making tax-aware decision-making critical for long-horizon investment strategies.

This project aims to build a fully automated system capable of long-horizon stock return prediction and tax-aware portfolio re-balancing for Mutual Funds constituents. The input to the predictive models consists of historical time-series data for each listed stock from 2010 onward, including OHLC prices, trading volume, volatility measures, and selected macroeconomic indicators. Using models designed for temporal dependencies—such as recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs)—the system generates return forecasts over 6–36-month horizons. These forecasts serve as inputs to a custom optimization module that recommends a re-balancing strategy accounting for both expected returns and the tax cost of realizing capital gains. The optimization component receives three inputs: predicted returns for each ETF constituent, the investor's current holdings and cost basis, and applicable capital-gains tax rules. It outputs an updated portfolio allocation and computes the resulting after-tax return relative to traditional Mutual Fund strategies. The overarching goal is to demonstrate that a data-driven, tax-aware forecasting and optimization framework can outperform conventional Mutual Fund's on an after-tax basis by explicitly modeling long-term trends and capital-gains drag.

2 Related Work

Work such as Moehle et al. (2021) demonstrated that incorporating capital-gains taxes directly into the allocation process can significantly improve after-tax outcomes, motivating the explicit tax modeling used in our custom tax-penalizing optimizer Moehle et al. (2021). In parallel, supervised learning approaches—such as those explored by Piovezan and de Andrade (2022)—show that linear regression and XGBoost models can provide competitive performance on ETF and stock time-series data Piovezan and de Andrade Junior (2022). Garcia et al. (2020) examines multi-stage portfolio optimization incorporating realistic transaction costs helping formulate how the re-balancing strategies can be executed with a penalty. Garcia et al. (2020) Complementing these traditional models, deep sequence architectures such as RNNs Elman (1990) and LSTMs Hochreiter and Schmidhuber (1997) have proven effective at capturing temporal dependencies, making them especially well suited for long-horizon prediction tasks relevant to our 6–36-month objectives.

A related stream of work applies reinforcement learning to portfolio management, most notably Jiang et al. (2017), who learn trading policies directly from historical price trajectories Jiang et al. (2017). Despite progress in forecasting and optimization separately, few studies integrate long-horizon predictive modeling with explicit tax-aware rebalancing. In practice, both mutual fund and

ETF rebalancing remain largely heuristic and advisor-driven. This project builds on prior research by combining robust ML forecasting techniques with a tax-aware optimization module, aiming to address a gap where academic work and industry practice currently diverge.

3 Dataset and Features

This project utilizes two primary data sources: (1) daily U.S. equity price data from January 2015 to June 2025, and (2) monthly mutual fund constituent data for selected sector-focused funds. This information was retrieved from publicly available financial tools (e.g., Yahoo Finance[1]. The dataset consisted of standard OHLCV attributes—Open, High, Low, Close, Adjusted Close, and Volume—for all stocks appearing in the selected funds on a per-day basis. The mutual fund dataset was sourced from the web-page available for each of the selected funds. This provides monthly holdings for each fund, including constituent stocks and their portfolio representation. For this study, two key investment sectors were focused on: Technology and Financial Services and the dataset was created for all stocks belonging to these sectors. The predictive features were derived from daily stock data, following widely adopted practices in quantitative finance. These engineered features include standardized values of the rolling averages (5-, 20-, 60-, 120-day windows), rolling volatility, lagged returns, momentum indicators (e.g., ROC, MACD), and relative price ratios. The selected features were designed to avoid any look ahead bias. Upon finalizing the features, the dataset was split chronologically into testing and training to avoid data leakage.

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

Dataset Component	**Frequency**	**Description**
Daily Stock Prices	Daily (2015–2025)	OHLCV data for all stocks appearing in the selected mutual funds
Mutual Fund Constituents	Monthly	Monthly list of fund holdings, including constituent tickers and portfolio weights
Sector Coverage	—	Technology and Financial Services sectors, selecting the top 5 mutual funds from each category
Engineered Features	Daily	Rolling means, rolling volatility, momentum indicators, lagged returns, and relative price ratios

Table 1: Overview of datasets used in the forecasting and optimization pipeline.

4 Methods

This project's methodology is organized into four sequential stages: (1) long-horizon stock price prediction, (2) tax-penalizing modeling, (3) mutual fund re-balancing optimization, and (4) simulation and evaluation. Together, these components form the end-to-end forecasting and tax-aware optimization system.

4.1 Stage 1: Long-Horizon Stock Price Prediction

The first stage aims to predict the 6-month forward price of individual stocks belonging to a specific sector in order to reduce unwanted variation from broad market regime changes and cross-sector heterogeneity. This restriction ensures that the forecasting models learn sector-specific price behavior rather than cross-sector signals that might be noisy.

This stage is formulated as a regression problem where given the engineered feature set $x_t \in \mathbb{R}^d$ for each stock at time t, the prediction target is the future price after a horizon of H trading days:

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

[1] https://finance.yahoo.com

Models like Linear Regression, XGBoost Regression, Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks were all evaluated. For RNN and LSTM models, predictions are computed as

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

where f_θ is a sequence model considering 'k' previous observations

The best-performing architecture for each sector was selected based on it's performance on the demarcated test set using the MSE and RMSE loss function's.

A dedicated predictive model is trained for each stock to generate long-horizon forecasts for it.

4.2 Stage 2: Tax-Penalizing Strategy

In order to incorporate the effects of taxation into the re-balancing process, capital gains is modeled at a lot level. For each asset i and its lot j, let

- $q_{i,j}$: 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 for that lot (short-term or long-term).

Realized gains from selling $\Delta q_{i,j}$ shares is

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

Tax owed on that lot is

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

Total tax penalty incurred due to re-balancing is

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

4.3 Stage 3: Mutual Fund re-balancing Optimization strategy

Using the forecasted returns and tax penalties, it was computed whether re-balancing a funds holdings improves its expected after-tax performance. Let w_i^{old} denote the current weight of asset i, w_i^{new} the proposed weight after re-balancing, and $\hat{r}_i$ the predicted H-month return.

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:

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

re-balancing is conducted only if:

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

4.4 Stage 4: Simulation and Evaluation

In order to evaluate the systems performance, a baseline representing each funds natural after-tax return was computed. The baseline metric is directly compared with the proposed tax-aware system in order to quantify the realized gains from the tax-aware system for the top 5 funds in a given sector. The baseline is formed as:

1. Identify the top ten holdings by weight as of November 2024 for the selected funds
2. Computing the after-tax portfolio value assuming an investor enters in November 2024 and follows the funds own monthly re-balancing schedule.

Using the same initial holdings, our system evaluates the potential improvement by:

1. Predicting H-month stock returns using the best model from Stage 1.
2. Computing and applying the tax-aware re-balancing rule from Stage 3.
3. Calculating realized gains, taxes, and net after-tax portfolio value.

The final metric quantifies the percentage improvement in after-tax portfolio value achieved by the proposed system relative to the baseline fund strategy.

5 Results

5.1 Long-horizon prediction experiments and results

Traditional machine Learning approaches like Linear Regression and XGBoost were initially trained for short-horizon prediction tasks (1 day to 1 week). During this phase Linear Regression produced low RMSE on held-out test data as compared to the XGBoost models, indicating that very near-term price movements contain largely linear structure. However, upon extending the target horizon to 6 and 12 months, Linear Regression exhibited strong underfitting and substantially higher RMSE; XGBoost yielded better robustness but still degraded at long horizons. (RMSE comparison plots for Linear Regression and XGBoost over 6-month and 12-month horizon windows). Due to the underfitting nature of the previous models, more complex model architectures were explored. To capture longer-range temporal structure, sequential models and deep learning techniques were explored. Upon performing a study the LSTM model architecture was identified as an ideal candidate. The models were trained with Adam (learning rate 1 x 10^{-3}), batch size 64, and MSE loss; these hyperparameters were chosen after a focused grid search on a validation set. Upon experimentation it was identified that tuning the sequence length k showed that shorter sequences of data outperformed longer windows, implying that providing the most relevant recent context improves long-horizon stability while excessively long histories introduced noise and caused overfitting. Overall, LSTMs produced the lowest RMSE and the most stable long-horizon forecasts.

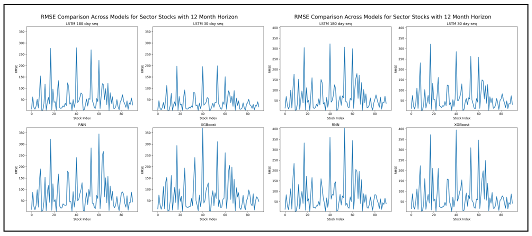

Figure 1: RMSE comparison of various models for 6-month and 12-month horizon windows

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 2: Avg RMSE for each stock by models

5.2 Results on Realised Gains Post Taxation

Once the most effective long-horizon prediction model was identified, the next objective was to use the evaluation framework to quantitatively prove whether a tax-aware re-balancing strategy could outperform leading mutual funds on an after-tax basis. This evaluation was performed on the top 10 funds, 5 from the Technology sector and 5 from the Financial Services sector, based on their returns in 2025. For each fund, its monthly constituent weights were collected and their natural re-balancing behavior was simulated. Using these allocations, we computed the baseline after-tax portfolio value by applying capital-gains taxation to each monthly re-balance. This baseline represents the realized gains and resulting tax liability that an investor would incur by strictly following each funds holdings schedule.

The proposed tax-aware system was then applied to the same initial constituents, and at each monthly step the long-horizon predictions to determine whether re-balancing was expected to improve after-tax value. A comparison table is provided here, summarizing the performance for each fund.

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 3: Net Returns Increase by tax-aware re-balancing system for Funds of Technolgy 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 4: Net Returns Increase by tax-aware re-balancing system for Funds of Financial sector

The net returns reported in the tables represent the difference in profits before and after taxation for the proposed system. Across the ten funds, we observed that returns before tax were negative for most cases, reflecting the uplift introduced by frequent re-balancing in the baseline strategies. For the two funds where pre-tax returns were positive, the long-horizon prediction model successfully recommended re-balancing decisions that outperformed the baseline approach. The proposed tax-aware system delivered higher after-tax returns for eight out of ten funds thus optimizing better for long-term strategies whilst also making necessary re-distributions when needed primarily by reducing the number of re-balancing events. Across the 10 funds, a net uplift of 1.21% is realized.

6 Conclusion / Future Work

This project developed an automated framework for long-horizon stock return forecasting and tax-aware mutual fund re-balancing. Among the models evaluated, LSTMs delivered the strongest predictive performance (based on RMSE scores of Table 2) due to their ability to model temporal structure better in financial time-series data. They were able to consistently beat out the RNN's and the XGBoost models across both the 6 and 12 month horizon. The results also indicate that the 6-month forecasts were consistently more accurate than the 12-month forecasts, suggesting that longer horizons require higher model capacity, larger datasets, or additional feature engineering (introduction of other market variables other than just the stock price) to better predict for longer horizons as required. The proposed tax-aware re-balancing strategy was able to derive meaningful gains in realized after-tax performance for a short window of time indicating that there is scope for optimization in this facet of the industry. There were cases where the re-balancing suggested by the system was better than the pre-tax returns of the existing Fund showing vast scope for improvement.

Further extensions could explore more advanced sequential models or even transformer-based architectures to further improve long-horizon predictability. Expanding the system to operate across multiple sectors would also require introducing contextual and macroeconomic features into the dataset to model cross-sector dependencies more effectively. Beyond the forecasting enhancements, the optimization layer can be made more realistic by allowing individual-specific tax configurations—such as income-based tax brackets, tax-exemption handling, or long-term holding benefits—to provide bespoke solutions based on the investors profile. Finally, introducing an exploration component alongside exploitation can enable the system to discover and evaluate previously unselected but potentially high-reward stocks, which can drastically improve benefits and maybe even tap into foreign markets.

References

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179–211.

Garcia, D., Garan, N., and Singh, A. (2020). Multi-period portfolio optimization with realistic transaction costs and constraints. *Journal of Financial Optimization*, 6(1):45–72.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jiang, Z., Xu, D., and Liang, J. (2017). Deep reinforcement learning for portfolio management. *arXiv preprint arXiv:1706.10059*.

Moehle, N., Kochenderfer, M. J., Boyd, S., and Ang, A. (2021). Tax-aware portfolio construction via convex optimization. *arXiv preprint arXiv:2008.04985*.

Piovezan, R. P. B. and de Andrade Junior, P. P. (2022). Machine learning method for return direction forecasting of exchange traded funds using classification and regression models. *arXiv preprint arXiv:2205.12746*.