On the predictability of ETF returns with technical predictors

April 5, 2024

Abstract

We use technical indicators, which are traditionally applied to stocks, to explore the predictability of equity ETFs in a random forest classification model. Our analysis suggests that technical signals constructed from stocks can forecast ETFs' future performance. We find that, without risk adjustment, equity ETF long-short portfolios achieve a monthly mean return of up to 0.76%, with a t-statistic of 2.75. In line with limits to arbitrage, we find that the level of market efficiency influences the predictability of technical indicators. It is evidenced that Chinese-focused ETFs outperform US-focused ones when applying the prediction model. Moreover, our findings are robust to different model settings.

Keywords: ETF, Predictability, Random Forest, Technical Analysis

1 Introduction

Technical analysis is based on the idea that historical price and volume data can be used to forecast future returns. It is a preferred tool among market participants due to its ease of understanding and implementation, clear entry signals, applicability in many markets, and fewer data requirements.[1] Menkhoff (2010) conducts a comprehensive survey of 692 fund managers across five countries, including the U.S., and their findings reveal that while technical analysis is commonly used in conjunction with fundamental analysis, its importance as a forecasting tool takes the lead, particularly when the focus shifts to short-term horizons. In addition, Nti et al. (2020) show that 66% of documents reviewed, which include published journal articles, conference proceedings papers, doctoral dissertations or supplementary unpublished academic working papers and reports, were based on technical analysis over 11 years (2007–2018). In this paper, we want to answer the question of whether technical trading signals, when applied using sophisticated machine learning methods, can predict the future performance of international exchange-traded funds (ETFs).

Exchange-traded funds (ETFs), often more liquid instruments than stocks, have seen substantial growth in both number and assets under management (AUM) in recent years, and the growth rate is nearly three times higher than that observed in traditional mutual funds over the period 2016-2022.[2] ETFs have diversified significantly and can be categorized based on their underlying assets into various types, including equity ETFs (e.g. SPDR S&P 500 ETF Trust), bond ETFs (e.g. VanEck Fallen Angel High Yield Bond ETF), commodity ETFs (e.g. Invesco DB Commodity Index Tracking Fund), and more. In this paper, we focus on equity ETFs, the dominant type in the ETF market.

Equity ETFs are investment funds traded on stock exchanges and track the performance of specific stocks, but there is a significant difference in the availability of historical data between ETFs and stocks. This data discrepancy affects the depth and breadth of potential empirical analysis for the ETF market, with stock markets having a longer historical record compared to ETFs, which are relatively new financial innovations. Individual stocks have extensive historical data going back decades, with records reaching back to the 1900s, providing valuable insights into market dynamics. In contrast, first appearing in the 1990s, ETFs, as newer entrants to the financial market, have a shorter history, which makes it difficult to do insightful time series analysis and especially out-of-sample analysis. While

[1] https://medium.com/@pannipa/advantages-and-disadvantages-of-technical-analysis-39c7c8b9a3ef.

[2] https://www.oliverwyman.com/our-expertise/insights/2023/may/exchange-traded-funds-are-fueling-market-opportunities.html.

the number of ETFs continued to grow in recent years, which enriches potential research subjects, this expansion does not equate to an extended historical record. This difference in data availability presents unique challenges in analyzing ETF returns and developing robust predictive models. Since ETFs can be seen as derivatives of stocks, and stocks can be seen as underlyings of equity ETFs, our study addresses this issue by using stock technical predictors to forecast ETF performance, creatively adapting to the historical data constraints of ETFs.

Our stock sample ranges from 1981 until 2021, and ETF sample spans from 2005 until 2022. All data are obtained from Datastream. Constructing various technical indicators as predictors using international individual stock and ETF data, we explore the predictive power using sophisticated machine learning methods for international ETFs. More specifically, we use technical indicators that have been developed for stocks and apply these to the new-established ETF market. Our objectives are twofold: firstly, to solve the data availability problem stemming from the ETF market's relatively short history, which poses challenges for conducting out-of-sample tests, we use stock data to train the models and make predictions in ETF data. Secondly, we conduct an international analysis to gain a deeper understanding of the predictability of ETFs focused on different countries.

We employ the tree-based model in our analysis, and the motivation stems from recent studies (e.g. Gu et al. (2020), Avramov et al. (2023), Leippold et al. (2022)). We explore the predictive performance of tree models for forecasting future individual international ETFs' performance using the global stock-trained model. Tree-based models are traditionally used for classification problems, and the experimental research conducted by Nabipour et al. (2020) reveals a substantial performance improvement in models when opting for binary data over continuous data. Thus, we employ the classification-based model which is comparable to Breitung (2023). Moreover, our study's objective is not to predict the exact return, typically addressed through a regression, but rather to establish a ranking of ETFs for each month. This ranking, crucial for long-short investment strategy, is obtained through a classification model, providing outperformance probabilities that guide investment decisions. Specifically, based on this ranking, we invest in the top 10% ETFs and short the bottom 10% ETFs for each month. Among all these tree-based models, we choose random forests, known for their robustness, which have been applied to predict stock prices, assess credit risk, and optimize investment portfolios. Ballings et al. (2015) set out to benchmark the performance of ensemble methods (Random Forest, AdaBoost, and Kernel Factory) against single classifier models (Neural Networks, Logistic Regression, Support Vector Machines, and K-Nearest Neighbors) in predicting stock price direction. It is determined that random forest emerged as the top performer in this analysis.

Our results reveal that technical indicators for stocks are useful for predicting the future ETF outperformance probability, which indicates a violation of the weak form of the efficient market hypothesis. We generate the long-short mean return of 0.76% and after adjusting for various risk factors the mean return is above 0.60%. Moreover, in line with previous research that highlights the short-term efficacy of technical signals, we observe that such predictability is most pronounced in the early months as we increase the rebalance interval from one month to 36 months. After categorizing the technical indicators into five groups, we find that models trained on a single category of indicators cannot generate a long-short return as high as the model trained on all indicators combined. Given the efficiency of the stock market that ETFs track, we find that the greater the efficiency of a market, the more challenging it becomes to predict future performance. The long-short return generated from the ETFs that invest in China is 0.87%, but only 0.48% in the US.

Accurately predicting future asset performance is not only crucial for investors and analysts to make informed investment decisions but also an major topic in academic research. Although the topic is compelling, there's no definitive guide on how to choose the correct models and what are the best predictors. To predict the ETFs' returns, a wide range of models, from traditional linear regression (e.g. Rompotis, 2011; Brown et al., 2021) to the latest machine learning and deep learning techniques (e.g. Liew and Mayster, 2017; Day and Lin, 2019), have been suggested. Rompotis (2011) examines predictability by regressing the ETFs' raw return on the four dummy variables representing the ETF star ratings, where each variable indicates whether an ETF belongs to a specific class (four stars through one star, respectively), with class-5 ETFs serving as the reference group through the model's constant term. Alternatively, Chen and Kuo (2006) propose a neural network-based decision support system to provide investors with suggestions on transaction timing and transaction strategies of Taiwan 50 index exchange-traded funds. Zhong and Enke (2019) predict the daily return direction of the SPDR S&P 500 ETF using a deliberately designed classification mining procedure based on hybrid machine learning algorithms. Machine learning techniques, as the latest advancements, are capable of dealing with complexity and non-linearity in data, and are especially suitable for large datasets.

Meanwhile, proposed predictors fall into two main categories: technical and non-technical indicators, each offering different insights for forecasting returns. Chen and Kuo (2006) employ various technical indicators, e.g. moving average, relative strength index, and stochastic oscillation, to construct ETF transaction strategies in Taiwan. Including three U.S. ETFs that track three indices of U.S. growth markets and six emerging market ETFs that track MSCI country index, Hsu et al. (2010) examine the predictive ability of technical trading rules: moving averages (MA) rules and filter rules (FR) and find with good executions and low transaction costs, the technical analysts in large institutions may be able to make profits in excess of risk premiums in the ETF market. Huang and Huang (2020) implement moving-average (MA) trading strategies on the ETF market and backtests the strategies against the buy-and-hold benchmark. They find while MA strategies generate positive average returns, they are lower than those of the buy-and-hold strategy. On the other hand, non-technical indicators for ETFs refer to factors not rooted in historical price and volume data, for example, fund flow and market sentiment. Brown et al. (2021) find that non-fundamental demand shocks, identified through ETF flows, can predict future ETF returns. Lee et al. (2021) use the data of 47 single-country exchange-traded funds traded in the U.S. from 36 countries during 2004–2017 and examines the impact of investor attention proxied by Google Search Volume Index and home country-specific factors on ETF returns. However, there is no clear evidence of whether the technical indicators are superior to the non-technical predictors or vice versa.

While extensive literature explores ETF return predictability, the majority focuses on particular markets or a select number of ETFs. For instance, Yang et al. (2010); Rompotis (2011); Zhong and Enke (2019); Brown et al. (2021); Chen (2023) concentrate on U.S. ETF return predictability. Day and Lin (2019) develop a robo-advisor with different machine learning and deep learning forecasting methodologies in the Taiwan ETF market. Jares and Lavin (2004) investigate foreign ETFs (Japan and Hong Kong) that trade on U.S. exchanges but provide broad exposure to foreign markets. Research on international ETFs yields relatively scarce results. Broman (2020) investigates the return predictability for 4560 twin-pairs of Exchange-Traded Funds (ETFs) from 15 country pairs. Zhang et al. (2023) use 29 country-/region-specific ETFs (including U.S.). With the new predictor, risk-neutral moments of returns, the authors predict future excess returns of country-/region-specific ETFs. Lee and Chen (2020) use 132 country-specific ETFs from 45 countries and examines whether social media (Twitter) happiness sentiment and country-level happiness sentiment indices predict cross-border ETF returns. Our empirical study distinguishes itself by compiling an extensive dataset of 9255 equity ETFs from 52 countries, constituting the most comprehensive dataset in this field to our knowledge.

Our study makes two significant contributions to this literature. First, we analyze the international equity ETFs with the random forest classification model. Unlike previous research that often focuses on a single market (e.g. US ETF market, Chinese ETF market) or specific ETFs (e.g. energy ETFs, gold ETFs), our analysis encompasses a broader perspective, considering international markets. This international perspective allows us to derive insights and conclusions that offer a more holistic view of the equity ETF landscape.

Second, we introduce an innovative approach by training our machine learning model on stock market data and subsequently applying it to ETFs, thereby addressing the challenge posed by the short historical data availability for ETFs. By employing this technique, we illuminate the potential for leveraging historical stock market data to make more informed predictions and investment decisions within the ETF market. Furthermore, our study diverges from conventional return prediction approaches by focusing on the relative ranking of ETFs. This shift allows us to better evaluate their potential for outperformance relative to one another. This emphasis on predicting outperformance probabilities, rather than absolute values, enhances the precision of our assessment of cross-sectional predictability and is less time-consuming when a large dataset is applied.

The remainder of the paper is structured as follows: Section 2 presents the data, machine learning methodology, and trading strategy. Section 3 explores the main results, while Section 4 conducts robustness checks. Finally, Section 5 offers the paper's concluding remarks.

2 Data and Methodology

2.1 Train-Test-validation Split

In the context of machine learning algorithms, the “training” subsample is used for model estimation, the “validation” subsample guides hyperparameter tuning based on forecast accuracy, and the “testing” subsample evaluates predictive performance. Besides the training phase, it is necessary to validate the model on unseen data to assess the performance and generalization capability of the model. This splitting framework is essential due to the need for rigorous model validation when working with historical stock and ETF data spanning several decades. Our analysis involves the use of two specific model settings, each adhering to a well-defined train-validate-test sequence. The first model is trained on stock data from 1981 to 2004, subsequently validated on ETF data from 2005 to 2010, and finally tested on

ETF data from 2011 to 2022. The second model is exclusively trained on ETF data from 2005 to 2010, validated on ETF data from 2011 to 2015, and tested on ETF data from 2016 to 2022. (see Figure 1) Our approach primarily centers on the first model, which emphasizes predictability across stocks and ETFs. This model is trained using global stock market data, and we call it the global stock-trained model. Subsequently, it is applied to the ETF market. Figure 1 displays the train-validation-test split across various scenarios and periods.

The training data set for stocks (ETF) encompasses a total of 2,474,549 (46,110) observations, providing a substantial volume of data for model training. When the model is trained on the stock market, the ETF validation and ETF test data sets consist of 46,110 and 527,243 observations, enabling a comprehensive assessment of the model's performance. Additionally, when the model is trained on the ETF market, the ETF validation and ETF test data sets comprise 127,825 and 399,418 observations.

[here insert Figure 1]

2.2 Stock and ETF Market Data

We collect international daily individual stock and equity ETF data across different exchanges from Datastream, including variables such as open price, high price, low price, closing price, turnover by volume, and return index. The daily stock and ETF data are utilized to calculate various monthly technical indicators and monthly returns. To be precise, the final stock data set (ETF data set) comprises monthly observations of individual stock returns (ETF returns) for the period from February 1981 to June 2021 (from January 2005 to December 2022). The stock returns (ETF returns) are calculated as the monthly percentage changes in the total return index (Datastream code: RI). The choice to begin the equity ETF data set from 2005 is attributed to the relatively limited availability of ETFs before that period, ensuring a comprehensive collection of ETFs for analysis.

Table 1 presents the summary statistics of the data set. It provides an overview of the number of observations, number of securities, number of countries, mean, and standard deviation for different subsets. In Panel A, the overall stock return data combines the training, validation, and testing data sets, resulting in a total of 9,181,979 observations. It includes 66,905 securities from 67 countries, with a monthly mean stock return of 1.36% and a standard deviation of 21% from the year 1981 to 2021. Moving on to the equity ETF data, the total data consists of 573,353 observations with 9,255 securities from 52 countries. The mean for this data is 0.53% with a standard deviation of 7% from January 2005 to December 2022. Panel B is structured as a train-validate-test split for the data set, segregating data into distinct periods to facilitate rigorous evaluation and validation of models. This panel encompasses four models as described in Figure 1. The training, validation, and testing periods for stocks and ETFs are distinctly outlined. The mean returns for stock training data is 1.33%, with a standard deviation of 21%, while the mean returns for ETF training data is 0.74%, with a standard deviation of 10%. The mean returns of ETF validation (test) data are 0.74% (0.51%) and 0.49% (0.52%) for two training models. Additionally, the average age of ETFs varies between 7 to 16 months, with average fees reported between 0.43% and 0.47%.

[here insert Table 1]

Both stock and ETF data sets are carefully collected and cleaned to ensure data accuracy and consistency. Any missing values are addressed by fill-forward technique, ensuring that no important information is lost. This technique uses the previous valid data point and serves as a proxy for subsequent missing values until new valid values are identified. Moreover, outliers are carefully identified and treated through winsorization at 0.1% level. To overcome survivorship bias, the data set encompasses both active and inactive ETFs and stocks.

2.3 Construction of Technical Indicators

To initiate our analysis, we employ the Technical Analysis Library (TA) in Python to work with the historical daily price and volume data of the stocks and ETFs under consideration.[3] TA is a popular open-source library that provides a wide range of technical analysis functions and indicators using time series data. Leveraging the capabilities of TA, we utilize our comprehensive data set to calculate various technical analysis indicators, for example, average directional movement index (ADX), relative strength index (RSI), awesome oscillator (AO), accumulation/distribution Index (ADI), and so on. These indicators are organized into five distinct categories: momentum indicators, trend indicators, volatility indicators, volume indicators, and other indicators. By computing these indicators, we are able to identify significant patterns, discern prevailing trends, and detect potential trading signals within the stock and ETF market. A comprehensive overview of all 172 technical indicators, categorized into five categories, can be found in Table A.1 in the Appendix.

[3] https://technical-analysis-library-in-python.readthedocs.io/en/latest/.

We construct the technical indicators on a daily level, but our investment strategy is rebalancing portfolios at a monthly level, therefore, to ensure our analysis does not incorporate information that is not accessible at the start of a month, we only include the most recent indicators available from the previous month. By focusing on these month-end indicators, we ensure consistency and capture the overall market trends and behavior at regular intervals. These technical indicators are then combined with the next month's returns, enabling us to examine the monthly cross-sectional predictability between technical indicators and the subsequent performance of ETFs.

2.4 Forecasting Methods and Portfolio Construction

We start with a crucial preliminary step — hyperparameter sensitivity analysis using the training and validation dataset. Hyperparameters, which control model complexity, are pivotal in machine learning but often lack clear theoretical guidance for optimal values and this analysis aims to assess how varying hyperparameters affect our model performance. We run different models using training data with different pairs of hyperparameters and then use the validation data to test the trained models' performances based on the accuracy metrics. Finally, we select the hyperparameters that generate the highest accuracy and choose that model for our further analysis. Notably, our choice of hyperparameters for the random forest model is informed by the research conducted by Gu et al. (2020). The specific parameter selections can be found in Table A.5 of their study, serving as a reference for our modeling approach. For detailed insights into our sample splitting scheme and hyperparameter tuning procedures for each model, please refer to Appendix Table A.2. We train predictive models using international stock and ETF data sets separately, and then we execute validation phases to get the best hyperparameters, lastly, we test the best-performing model interactively across both asset classes.

Then, leveraging the random forest classification model, we effectively assess the outperformance probabilities of ETFs at the monthly level. The model allows us to determine the probability of ETFs exceeding the performance of the other ETFs. For each month, we label the data as '0' if the return falls below the median return of that month, and as '1' otherwise. During the prediction phase, the individual trees in the random forest classification model independently generate their predictions for each class (0, 1), and the probability of a specific class (here we focus on the class labeled as '1' which represents the outperforming category) is obtained by averaging the probabilities assigned to that class by all the trees in the forest. The random forest classification model serves as our baseline model, providing valuable insights into the relative rankings of ETFs.

Besides random forest classification models, we use a logistic regression to estimate the probability of an event occurring based on input features. We also incorporate other machine learning techniques for robustness checks, including decision tree classification (DT), which is a supervised learning model that partitions the feature space into segments, using a tree-like structure, to make categorical predictions based on the input features; gradient boosted decision trees (GBDT), which is an ensemble learning model that combines multiple decision trees, sequentially trained to correct the errors made by the previous trees, resulting in a strong predictive model capable of handling complex relationships in the data, and XGBoost classification, which is an optimized implementation of gradient boosting that incorporates regularization techniques to prevent overfitting, making it a powerful and accurate model for classification tasks.

In addition to our primary focus on ranking ETFs, we also employ a diverse set of regression models to further validate our results. These models include linear regression, decision trees regression (DT), gradient boosting regression trees (GBRT), and XGBoost regression. While each model has its unique characteristics, our objective remains consistent: to enhance the robustness of our analysis by exploring various regression techniques. It's important to note that despite potential variations in training times, the precise return values are not crucial for our portfolio construction approach, which aligns with our ultimate goal.

From a modeling perspective, tree-based models, like random forest models, have a limitation when it comes to capturing and representing time trends in panel data. In our data set, we collect observations over multiple periods for each cross-sectional unit, which means we have panel data. Time trends are one aspect of our data that we can not ignore, but these tree models are primarily designed to capture cross-sectional variation, where different stocks and ETFs exhibit distinct behaviors. Although our analysis involves a monthly cross-sectional ranking prediction, it's crucial to note that we incorporate a time series perspective when generating technical indicators. Technical analysis predominantly relies on time-series data, emphasizing temporal patterns over cross-sectional predictability. By incorporating these technical indicators, we effectively incorporate the time dimension into our model, allowing us to gain insights into the time trends integral to our analysis.

Based on the predicted outperformance probabilities obtained from the random forest classification models, we construct long-short portfolios. We assume that investors rebalance their portfolios monthly. For each month, we sort the available ETFs based on their respective predicted outperformance probabilities using the previous technical indicators. We implement a zero-investment strategy, which involves taking long positions on the top 10% of ETFs exhibiting the highest outperformance probability and short positions on the bottom 10% with the lowest probability. The portfolios are constructed using equal weights for each ETF position. We apply the same methods for constructing the stock portfolios.

3 Empirical Results

This section provides an overview of our empirical findings. In subsections 3.1 and 3.2, we mainly explore the predictive power of stock technical indicators on ETFs. Moving on to subsection 3.3, we narrow our focus to a specific subset of technical indicators. We examine whether the predictability is more pronounced for specific indicator categories. In subsections 3.4, we investigate ETFs that focus on various investment areas. Our aim is to determine if the ETFs that are invested in less efficient markets exhibit greater predictability compared to the ETFs that are invested in more efficient markets. In subsection 3.6, we describe the characteristics of the long and short portfolios in terms of ETF fees and the age of the ETFs, helping us determine which ETFs with specific attributes are likely to be included in the long portfolio and which ones in the short portfolio.

3.1 Short-Term Predictability

Before running the random forest classification model, we need to choose the optimal hyperparameters using validation set. Figure 2 shows the results of our hyperparameter sensitivity analysis for the two specified model settings. On the x-axis, we display the maximum features incorporated into each model, while the y-axis represents the accuracy of the validation data. Each line on the graph corresponds to a different maximum depth level, and the red circles highlight the best-performing hyperparameters for each specific model setting. This analysis allows us to pinpoint the most suitable hyperparameter configurations for our models, enhancing their predictive capabilities. We focus on the model trained on stock data and applied to ETFs, and the result in the first sub-figure indicates that model accuracy improves with an increase in the maximum features. However, this improvement is limited when the maximum features reach 20. Our findings suggest that simply increasing the model's complexity—by increasing the maximum number of features and the model's depth—does not necessarily yield better performance. This in turn explains that hyperparameter sensitivity analysis is necessary. We summarize the optimal hyperparameters in Table 2. In consistent with Gu et al. (2020), we use 300 estimators for all the models. For the model trained on stocks and validated on ETFs, the optimal settings are a maximum depth of 6 and 30 maximum features. Conversely, for the model trained and validated on ETFs, the ideal parameters are a maximum depth of 5 and 50 maximum features.

[here insert Figure 2]

[here insert Table 2]

Moving to the next step, we proceed by training our model utilizing technical indicators extracted from the global stock and ETF market with the selected hyperparameters. Table 3 provides empirical insights into the predictive potential of technical signals sourced from both the global stock market and the global ETF market with monthly rebalancing. The left section of the table displays results obtained from models trained on stock data spanning the years 1981 to 2004. Then the model is validated using the data from 2005 to 2010 in the ETF market. This model is finally applied to evaluate the equity ETF market from 2011 onwards. Conversely, the right section of the table shows outcomes from models trained on ETF data spanning the years 2005 to 2010, validated from 2011 to 2015, and subsequently used to assess the equity ETF market from 2016 onwards.

[here insert Table 3]

In the case of the model trained on stock information, we observe a substantial monthly long-short mean return of 0.76% with high statistical significance (t-statistic: 2.75). Similarly, after adjusting the market risks, the model yields a long-short mean return of 0.61% with a t-statistic of 2.28. The risk-adjusted long-short mean returns after Fama–French three-factor model and Carhart four-factor model are 0.60% (t-statistic: 2.24) and 0.62% (t-statistic: 2.32) respectively. For the model trained on ETF technical indicators, we find a somewhat lower raw return of 0.58% with less statistical significance (t-statistic: 1.84. Additionally, in terms of risk-adjusted long-short mean return, the ETF-trained model also did not outperform the stock-trained model. It's important to consider that the evaluation periods for the two models are different. The evaluation period for the ETF-trained model is much less than for the stock-trained model due to the shorter history of the ETF market. These findings highlight the importance and effectiveness of using the stock-trained model.

In summary, the findings in Table 3 strongly support the concept that stock technical indicators from closely linked markets can effectively predict equity ETFs' performance. This underscores the interdependence and responsiveness of ETFs to relevant data from correlated markets, illuminating the relationship between stock and ETF markets. In the following analysis, we concentrate on our global stock-trained model.

In addition to adjusting for risk-adjusted returns in Table 3, we consider the size and style effects in our analysis of long and short portfolios. The adjustment is crucial because the long or short portfolios may predominantly comprise ETFs that invest in large-cap stocks or value stocks, potentially biasing the results. To neutralize the size and style effect, we collect the Morningstar equity style box which categorizes funds into nine distinct categories based on size (small, mid, and large) and style (growth, blend, and value). We adjust the returns by subtracting the mean returns of ETFs within each size and style category before calculating the overall portfolio return. The adjusted size-neutral and style-neutral returns are presented in Table 4. We observe that the style-neutral return (0.57%) is higher than the size-neutral return (0.49%) and with higher significance.

[here insert Table 4]

Moreover, Figure 3 depicts the out-of-sample cumulative return of the long-short portfolio in the ETF market from the year 2011 onwards, utilizing the model trained on the global stock dataset. The y-axis represents the cumulative returns, which range from 0 to 1.75%. The x-axis is time, marked at two-year intervals. It shows an overall upward-sloping curve in returns, starting near zero and peaking above 1.5% towards the end of the period. Notably, there is a sharp decline post-2014 and a sharp increase in returns after 2020, indicating periodic fluctuations.

[here insert Figure 3]

3.2 Long-Term Analysis

Employing the global stock-trained model, our research reveals that at the monthly frequency, stock indicators play a prominent role in forecasting ETF returns, as evident in Table 3. We next examine the long-term horizon and we increase the rebalance periods. In Figure 4, we illustrate the outcomes of portfolio rebalancing at different time intervals. Starting from the left, the first graph represents the rebalancing every 6 months, followed by the 12-month interval, then the 24-month interval, and finally, the 36-month interval on the far right. The y-axes measure monthly returns, fluctuating above and below zero. We find that, in the ETF markets, robust predictability is primarily confined to the first month and attenuates in the subsequent months. This phenomenon indicates the presence of predictability in the short term while underscoring the absence of predictability in the long term for ETF returns.

[here insert Figure 4]

One reason could be the creation and redemption process of ETFs. The involvement of Authorized Participants (APs) in this process contributes to the unique characteristics of ETFs. APs play a critical role in maintaining the structure and liquidity of ETFs by creating and redeeming ETF shares based on supply and demand dynamics. This creation-redemption mechanism ensures that the supply of ETF shares in the open market aligns with the corresponding demand, facilitating fair pricing of ETFs. As a result, the long-term predictability of ETF returns is expected to be absent, as the mechanism strives to keep ETF prices in line with the underlying assets they track. However, in the short term, there exists the possibility of predictability. This is due to the inherent characteristics of ETFs, such as tracking errors, which arise from discrepancies between the ETF's performance and the underlying stocks it aims to replicate. These tracking errors can create short-term opportunities for predictability in ETF prices.

3.3 Indicator Importance

To gain insights into the contribution of each feature category in the prediction of ETF ranking, we perform a ”Feature Importance” analysis, grouping the features into five categories as described in the Construction of Technical Indicators section. The grouped feature importance from the random forest classification model is as follows: With a relative importance of 0.36, the volatility group emphasizes the greatest significance among volatility-related indicators. Volatility measures, such as average true range or bollinger bands, contribute significantly to the model's ability to classify and predict outcomes. The trend group, with a relative importance of 0.25, ranks as the second most significant, following the volatility group. Trend indicators, such as exponential moving average or moving average convergence divergence, provide insights into the direction and strength of price movements. The momentum group of features has a relative importance of around 0.25 and exerts a comparable influence on the classification model as the trend group. It suggests that the recent price movements and trends of the securities are influential factors. The volume group has the lowest relative importance of around 0.02, indicating that volume-related indicators have a relatively minor influence on the classification model. However, they still contribute to some extent in assessing predictability. The other group has a relative importance of 0.11, indicating that other miscellaneous factors not explicitly categorized in the other groups contribute to the classification model too. (see Figure 5)

[here insert Figure 5]

However, the essence of "Feature Importance" analysis is the degree of dependency that a well-trained model has on features, and it does not represent the features' ability to generalize to unseen data (the test set). Especially when there is a distribution shift between the training and testing data sets (here in our study, the stock data distribution and the ETF data distribution are not the same), this default bias in the model's feature importance analysis can become more significant. These grouped feature importances in Figure 5 demonstrate the varying contributions of different categories of technical indicators in the random forest classification model only based on training data, with volatility and trend-related factors appearing to have the most substantial impact.

To assess feature contributions in the test data, we train separate models using each group of stock technical indicators and subsequently validate and test them in the ETF market. The empirical findings, presented in Table 5, examine the predictive power of stock technical indicators across the five categories and we find that not all categories are significant in making predictions. The table shows significant differences in ETF returns between long and short positions for momentum, volatility, and other indicators, which generate long-short returns of 0.66%, 0.64%, and 0.61%, respectively. The long-short returns for each category are smaller than the 0.76% achieved from the model that includes all categories. (see Table 3)

One thing that needs to be mentioned is that, after the hyperparameter sensitivity analysis with the set of hyperparameters from Gu et al. (2020), we end up with shallow and simplest models trained on trend and volume indicators. The optimal max depth from the validation data set is 1 and the models do not reliably predict the probability of outperformance. The predicted probabilities exhibit low variation, suggesting that the models are not confident enough to strongly favor one class over the other. This renders them ineffective for ranking purposes and for selecting long and short portfolios. There is a trade-off between the simplicity of a model, which enhances its ability to generalize to unseen data, and the model's confidence in its predictions, as indicated by the variability in predicted probabilities. While a simpler model may exhibit superior generalization capabilities, it might also fail to capture the data's complexity adequately, potentially leading to underfitting. Thus, we go beyond the hyperparameter's set from Gu et al. (2020). We use the same number of estimators as proposed by Gu et al. (2020) but increase the max depth and max features to increase the model complexity. We allow the trees to grow without constraints and consider many features for splits. Our objective is to balance model complexity with generalization ability, thereby obtaining more varied and meaningful predictions of outperformance probability. In the end, the trend and volume indicators demonstrate weaker performance, with returns of 0.29% and 0.04% respectively, and without statistical significance. The potential reason could be that the optimal model performance heavily relies on the quality and relevance of the features used. The current selection of features might not be providing enough discriminative information for the model.

[here insert Table 5]

To confirm our analysis that the trend and volume indicators generate weak performance, even with a complex machine learning model, we conduct a logistic regression, which is linear, for comparison. The results, detailed in Table 6, indicate that models trained on momentum indicators show a 0.58% long-short mean return with high significance. Conversely, models trained on volume indicators exhibit a negative long-short mean return at -0.28%. Although the model trained on trend indicators shows a positive long-short return, this is not statistically significant. These findings from the logistic regression are consistent with those from our deep random forest models with the max depth 115 for the trend model and 225 for the volume model, reinforcing the conclusion that 'volume' and 'trend' indicators, when used independently, may not possess strong predictive power. This suggests that the complexity added to the random forest models does not translate to increased forecasting reliability for these particular indicators and that the indicators may not have strong predictive power.

[here insert Table 6]

3.4 Different Investment Area

Utilizing the previous global stock-trained model, we conduct an analysis of predictability across various markets. We categorize the ETFs based on their geographic focus. "Geographic focus" refers to the specific regions or countries in which these ETFs invest, i.e. investment area, rather than their countries of origin or domicile. Our data set comprises information on the geographic focus of 7,639 ETFs sourced from Refinitiv. In Table A.3 we present the number of ETFs in each investment area. For analysis, we include only those areas that have more than 200 ETFs to ensure a substantial sample size. The majority of the ETFs in our data set invest in the US market.

Table 7 presents the evaluation results of ETF markets based on different geographic focuses. The evaluation includes ETFs that target global stocks, European stocks, US stocks, Chinese stocks, Korean stocks, and global stocks excluding US stocks. Using the global stock-trained model to make predictions on specific areas, we observe that all long-short ETF returns are statistically significant, although their magnitudes differ. Notably, ETFs focusing on the Chinese market achieve the highest long-short return at 0.87%, indicating a potentially favorable investment opportunity. On the other hand, ETFs concentrating on the US market exhibit the lowest long-short return at 0.48%. These findings suggest that the geographic focus of ETFs plays a significant role in their performance, with Chinese-focused ETFs demonstrating stronger potential for generating positive returns compared to US-focused ETFs. This observation aligns with the notion that the predictability of ETFs is influenced by the efficiency of the underlying stock market. In more efficient markets (e.g. US), where information is quickly incorporated into prices, predictability tends to be lower. Conversely, in less efficient markets (e.g. China), there may be greater opportunities for predictability and potentially higher long-short returns.

[here insert Table 7]

In addition to evaluating monthly predictability, we also examine the cumulative returns for each investment area over time. As depicted in Figure 6, initially, the cumulative returns for each investment area show little disparity. However, as time passes, ETFs investing in the Chinese market demonstrate significantly higher cumulative returns in comparison to ETFs focused on the US market. This observation underscores that the reduced predictability observed in more efficient markets not only affects short-term outcomes but also extends to long-term investment performance. The divergence in cumulative returns over time suggests that the Chinese market provides greater opportunities for generating sustained profitability compared to the US market. These results emphasize the importance of considering the efficiency of the underlying market when making investment decisions in ETFs.

[here insert Figure 6]

3.5 Other Characteristics of the Portfolios

In the subsequent analysis, we examine specific characteristics of the ETFs held within the long and short portfolios. The portfolios are constructed from our global stock-trained model prediction. We focus on two key characteristics: ETF age, measured in months since inception, and the net expense ratio, presented in percentage, and both are sourced from Morningstar.

Table 8 provides an overview of the aforementioned ETF characteristics, including both the long and short portfolios and the differences between long and short portfolios. On average, the ETFs within the long portfolios exhibit an age of approximately 64.78 months, while those in the short portfolios have an average age of 53.46 months. Notably, there is a significant age disparity between the two, the long portfolio is 11.32 months older than the short portfolio on average. Regarding fees, the long portfolio carries an average fee of 0.49%, while the short portfolio boasts a slightly higher fee of 0.60%, resulting in a substantial fee difference of -0.11% that holds high statistical significance (t-statistic: -7.20).

[here insert Table 8]

In the Fama-MacBeth analysis presented in Table 9, we examine the influence of ETF characteristics on their monthly returns, using the outperformance probability, age, and fees as independent variables, and incorporating year-fixed effects to account for time-related variations. This analysis reveals a significant positive relationship between ETF returns and outperformance probability (with a coefficient of 0.033 and a t-statistic of 1.72), which confirms the effectiveness of our global stock-trained model. Conversely, a negative correlation between fees and returns is consistent with our expectations and with the results in Table 8, indicating that higher fees detrimentally affect returns; specifically, a 1-basis point (bps) increase in fees (%) results in a 12-basis point decrease in monthly returns. While the analysis also suggests a positive trend between ETF age and returns, mirroring findings from Table 8, however, this relationship does not reach statistical significance. These findings shed light on the critical role that various characteristics play in determining the returns of these investment vehicles.

[here insert Table 9]

4 Robustness

4.1 Various models

In the previous section, our analysis primarily focuses on the implementation of the random forest classification model. Now we aim to broaden our analytical framework by incorporating a range of alternative models. This expansion includes not only different classification-based models such as Linear Models, Decision Trees (DT), Gradient Boosting Decision Trees (GBDT), and XGBoost but also regression-based models. By diversifying our modeling approach, we seek to enhance the comprehensiveness and resilience of our analysis. The results obtained from this extended experimentation, utilizing the same training and testing data as presented in Figure 1 (training data in the stock market, and testing data in either the stock market or ETF market), are outlined in Table 10.

Table 10 Panel A shows the evaluation of the stock market. The XGBoost model emerges as the standout performer among the classification models in the stock market, achieving the highest long-short return. This model delivers a long return of 1.83% and a short return of -0.05%, resulting in a long-short mean of 1.88%. This indicates its remarkable ability to capture profitable trading opportunities effectively. In Panel B, focusing on the ETF market, the XGBoost model continues to be the top performer. It generates a long return of 0.96% and a short return of 0.13%, resulting in a substantially high long-short mean of 0.83%, supported by high statistical significance. The outperformance of XGBoost compared to others is not surprising, because of its advanced features such as regularization, system optimization, parallel processing, and efficient handling of missing values, that enhance its efficiency, accuracy, and ability to control over-fitting. Notably, when comparing the linear model (Logit model) with other tree models, the return difference appears relatively modest. This suggests that the linear model is also proficient in this context, indicating its competency in generating favorable results.

On the right side of Table 10, regression models trained on stocks demonstrate much higher returns in the stock market but relatively lower returns in the ETF market compared with the classification models. This phenomenon finds support in a recent study by Breitung (2023), which underscores the propensity of regression models to place disproportionate emphasis on stocks with the most extreme return behaviors during their training process. Such overemphasis, coupled with the inherent challenge of regression models in effectively generalizing to new data, contributes to the observed performance discrepancy. As a result, we observe a reduction in long-short returns within the ETF market for regression models. The presence of '-' symbols in the table indicates the absence of results for the GBRT model. This exclusion is due to the GBRT model's extensive computational time, making its use impractical and infeasible for this study. It is important to note that regression-based models entail significantly longer execution times when compared to their classification-based counterparts. In our analysis, to implement the long-short trading strategy, our target is the ranking of the ETFs based on the outperformance probability, and the exact return prediction is not necessary. In light of this, we prioritize the use of classification-based models for our analysis.

[here insert Table 10]

4.2 Test on bond ETF market

Next, we collect bond ETF data from Datastream and apply the global stock-trained model to make predictions on the bond ETF market. Applying both regression-based and classification-based models using the stock technical indicators, it is apparent that these models are not well-suited for this particular financial domain. The inherent distinctions between global stocks and bond ETFs, where the underlying assets consist of corporate bonds and government bonds, significantly contribute to the limited effectiveness of these models when applied to the bond ETF market.

Table 11 presents the performance evaluation of various models when tested on the bond ETF market. In the "classification-based" category, it is apparent that the decision tree model exhibits the best performance with a relatively high long-short mean return of 0.40% and a t-statistic of 2.28. Besides, models such as the linear model, random forest, GBDT, and XGBoost display a range of performance levels, yielding long-short mean returns that span from -0.12% to 0.03% without statistical significance. These returns are notably lower when compared to the long-short returns previously generated in the equity ETF market (see Table 3 and Table 10). In the "regression-based" category, the decision tree model shows a low long-short mean return of 0.02% and no statistical significance. The random forest model demonstrates a somewhat stronger performance with a return of 0.18% but also not significant. These findings collectively suggest that the application of these models to the bond ETF market is challenging, as their predictive power is notably limited.

Overall, the findings indicate that relying on models initially trained on stock data does not lead to favorable outcomes when applied to the bond ETF market and it requires tailored modeling approaches to effectively capture its specific characteristics. It is suggested that models trained on corporate bond and government bond data, as opposed to stock data, might be more effective in the bond ETF market.

[here insert Table 11]

5 Conclusion

To conclude, by understanding the interconnected nature of stocks and equity ETFs and dealing with the data history shortage problem, our research finds that the technical predictors, which have originally been used for stock selection, effectively predict the ETFs' future performance. We observe the presence of outperformance predictability in the international ETF market for the short term using the global stock-trained model, suggesting that monthly trading strategies are successful. However, our findings do not support the same level of predictability in the long run, implying that long-term investment strategies may not yield comparable results in the ETF market. The feature importance analysis in the random forest classification model shows that volatility and trend indicators are crucial for model prediction based solely on stock training data. Using different models trained on grouped stock indicators, the out-of-sample tests on ETF testing data reveal that volatility and volume indicators significantly influence trading strategies with notable differences in returns between long and short positions. In addition, consistent with expectations, we identify that low-fee ETFs tend to be favored for investment, while high-fee ETFs are more frequently chosen for short positions. The net expense ratio is associated with statistically significant ETF returns. It is also important to note that the age of the ETFs does not explain the ETF returns. However, when constructing long and short portfolios, it is observed that older ETFs are preferred for buying, while younger ETFs are favored for selling. Finally, examining the stock market's underlying efficiency, our findings reveal that ETFs tracking more efficient stock markets exhibit lower predictability. This suggests that in markets where information is rapidly incorporated into stock prices, fewer opportunities exist to predict future price movements using historical data.

There are limitations to our research. Firstly, a substantial time gap exists between the training and testing data periods, as illustrated in Figure 1. Our global stock-trained model is trained on stock data spanning from 1981 to 2004 and subsequently tested on ETF market data from 2011 to 2022, with a distinct six-year validation period in between. This simple separation based on time points affects the model's generalization to new market conditions because it lacks information from the subsequent six years. Secondly, we recognize the dynamic nature of financial markets, known for their adaptability to external influences, for example in 2008 the global financial crisis, in 2020 the COVID-19 pandemic, and regulatory changes, which can result in structural shifts in market characteristics. Partitioning our data set into training, validation, and test sets, we recognize that such structural breaks can disrupt market relationships and potentially impact the model's predictive accuracy.

References

Avramov, D., Cheng, S., and Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, 69(5):2587–2619.

Ballings, M., Van den Poel, D., Hespeels, N., and Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert systems with Applications*, 42(20):7046–7056.

Breitung, C. (2023). Automated stock picking using random forests. *Journal of Empirical Finance*, 72:532–556.

Broman, M. S. (2020). Local demand shocks, excess comovement and return predictability. *Journal of Banking & Finance*, 119:105910.

Brown, D. C., Davies, S. W., and Ringgenberg, M. C. (2021). Etf arbitrage, non-fundamental demand, and return predictability. *Review of Finance*, 25(4):937–972.

Chen, C.-L. and Kuo, M.-H. (2006). An etf trading decision support system by using neural network and technical indicators. In *The 2006 IEEE international joint conference on neural network proceedings*, pages 2394–2401. IEEE.

Chen, J. (2023). Etf shorting activity, return predictability and information efficiency of underlying securities. *Available at SSRN 4527115*.

Day, M.-Y. and Lin, J.-T. (2019). Artificial intelligence for etf market prediction and portfolio optimization. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 1026–1033.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hsu, P.-H., Hsu, Y.-C., and Kuan, C.-M. (2010). Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. *Journal of Empirical Finance*, 17(3):471–484.

Huang, J.-Z. and Huang, Z. J. (2020). Testing moving average trading strategies on etfs. *Journal of Empirical Finance*, 57:16–32.

Jares, T. E. and Lavin, A. M. (2004). Japan and hong kong exchange-traded funds (etfs): Discounts, returns, and trading strategies. *Journal of Financial Services Research*, 25:57–

69.

Lee, C.-C. and Chen, M.-P. (2020). Happiness sentiments and the prediction of cross-border country exchange-traded fund returns. *The North American Journal of Economics and Finance*, 54:101254.

Lee, C.-C., Chen, M.-P., and Lee, C.-C. (2021). Investor attention, etf returns, and country-specific factors. *Research in International Business and Finance*, 56:101386.

Leippold, M., Wang, Q., and Zhou, W. (2022). Machine learning in the chinese stock market. *Journal of Financial Economics*, 145(2):64–82.

Liew, J. K.-S. and Mayster, B. (2017). Forecasting etfs with machine learning algorithms. *The Journal of Alternative Investments*, 20(3):58–78.

Menkhoff, L. (2010). The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, 34(11):2573–2586.

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., and Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *IEEE Access*, 8:150199–150212.

Nti, I. K., Adekoya, A. F., and Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, 53(4):3007–3057.

Rompotis, G. G. (2011). Predictable patterns in etfs' return and tracking error. *Studies in Economics and Finance*, 28(1):14–35.

Yang, J., Cabrera, J., and Wang, T. (2010). Nonlinearity, data-snooping, and stock index etf return predictability. *European Journal of Operational Research*, 200(2):498–507.

Zhang, J., Ruan, X., and Zhang, J. E. (2023). Risk-neutral moments and return predictability: International evidence. *Journal of Forecasting*, 42(5):1086–1111.

Zhong, X. and Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. *Financial Innovation*, 5(1):1–20.

Table 1: Summary Statistics

This table summarizes our dataset of both stocks and ETFs.We count the number of observations (# of Obs.), the number of stocks/ETFs (# of Securities), the number of countries that are covered in our dataset (# of Countries), the mean of stocks/ETFs, and standard deviation of stocks/ETFs. We include the average age and fee of ETFs.

	periods	# of Obs.	# of Securities	# of Countries	Mean	Std.	Ave. Age	Ave. Fee
Panel A: All Data								
Stock return	1981.02-2021.05	9,181,979	66,905	67	1.36%	0.21	-	-
ETF return	2005.01-2022.12	573,353	9,255	52	0.53%	0.07	7	0.43
Panel B: Train-Validate-Test Split Data								
Stock return train	1981.02-2004.12	2,474,549	34,527	52	1.33%	0.21	-	-
ETF return validate	2005.01-2010.12	46,110	1,300	29	0.74%	0.10	16	0.47
ETF return test	2011.01-2022.12	527,243	9,234	52	0.51%	0.06	7	0.43
ETF return train	2005.01-2010.12	46,110	1,300	29	0.74%	0.10	16	0.47
ETF return validate	2011.01-2015.12	127,825	3,279	42	0.49%	0.06	13	0.45
ETF return test	2016.01-2022.12	399,418	9,234	51	0.52%	0.06	7	0.43

Table 2: Model Parameters After Tuning

This table shows the best parameters that are tuned. We include the number of trees in the forest (n_estimators), the maximum depth of the tree (max_depth), and the number of features to consider when looking for the best split (max_features).

Train on Stocks			Train on ETFs		
n_estimators	max_depth	max_features	n_estimators	max_depth	max_features
300	6	30	300	5	50

Table 3: One-Month Predictability

This table reports cross-asset predictability results for two different training scenarios: one with models trained on stock data (1981-2004) and the other with models trained on ETF data (2005-2010). The evaluation periods start in 2011 for stock trained model and in 2016 for ETF trained model. The table is divided into four panels. Panel A shows the evaluation of these models when applied to ETF raw return predictions, displaying metrics such as long return, short return, long-short mean return, and mean return. Panel B, C, D present the evaluation results for risk adjusted ETF raw returns. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Train on Stocks			Train on ETFs		
long return	short return	l-s mean return	long return	short return	l-s mean return
Panel A: ETF raw return					
1.02%	0.26%	0.76%*** (2.75)	0.98%	0.40%	0.58%* (1.84)
Panel B: ETF CAPM Alpha					
0.43%	-0.23%	0.61%** (2.28)	0.30%	-0.14%	0.34% (1.17)
Panel C: ETF 3 Factor Alpha					
0.44%	-0.20%	0.60%** (2.24)	0.32%	-0.11%	0.35% (1.19)
Panel D: ETF 4 Facor Alpha					
0.56%	-0.11%	0.62%** (2.32)	0.42%	-0.18%	0.52%* (1.81)

Table 4: Adjusted Predictability

This table reports cross-asset predictability results for two different training scenarios: one with models trained on stock data (1981-2004) and the other with models trained on ETF data (2005-2010). The table is divided into two panels. Panel A shows the evaluation of these models when applied to stock market predictions, displaying metrics such as long return, short return, long-short mean return, and mean return. Panel B presents the corresponding evaluation results for ETFs. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

	long return	short return	l-s mean return
Panel A: Size Neutral ETF raw return			
	0.37%	-0.12%	0.49%* (1.74)
Panel B: Style Neutral ETF raw return			
	0.41%	-0.16%	0.57%** (2.14)

Table 5: Selected Indicators

This table reports the classification results when we restricted our training data to different categories of stock technical indicators. The models are trained from 1981 until 2004 and evaluated from 2011 until 2022. The results are tested on the ETF market. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Indicators	n_estimators	max_depth	max_features	long return	short return	l-s mean return
momentum	300	6	20	0.91%	0.26%	0.66%***
						(2.57)
trend	300	115	sqrt	0.72%	0.43%	0.29%
						(1.55)
volatility	300	3	sqrt	0.90%	0.26%	0.64%**
						(2.23)
volume	300	225	sqrt	0.64%	0.60%	0.04%
						(0.25)
others	300	4	3	0.80%	0.19%	0.61%**
						(2.15)

Table 6: Selected Indicators from Logistic Regression

This table reports the classification results when we restricted our training data to different categories of stock technical indicators. The models are trained from 1981 until 2004 and evaluated from 2011 until 2022. The results are tested on the ETF market. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

Indicators	long return	short return	l-s mean return
momentum	0.92%	0.34%	0.58%*** (2.98)
trend	0.67%	0.41%	0.26% (1.29)
volatility	0.80%	0.38%	0.42% (1.42)
volume	0.44%	0.72%	-0.28%** (-2.45)
others	0.83%	0.08%	0.75%*** (2.78)

Table 7: Model test for different investment area

This table reports the results from the global stock-trained model testing on the different investment areas from 2011 onwards. We show, for different investment areas, the long return, short return, and long-short return with statistical significance.

investment area	long ret	short ret	l-s mean
Global	1.07%	0.41%	0.66%** (2.35)
Europe	1.08%	0.35%	0.73%*** (2.60)
US	0.82%	0.34%	0.48%* (1.60)
China	1.07%	0.20%	0.87%*** (3.92)
Korea	1.05%	0.36%	0.69%** (2.44)
Global Ex US	1.08%	0.36%	0.72%** (2.57)

Table 8: The characteristic of the ETFs

This table presents information on the age of these ETFs and their net expense ratios. The age of ETFs, measured in months, is divided into categories for the long portfolio and short portfolio, along with the age difference between the two. Similarly, the net expense ratios are provided for both portfolios, with the percentage sign (%) omitted for brevity, and the fee difference is calculated. The table's statistical significance is denoted by asterisks and t-statistics in parentheses.

	Age of ETF			Net Expense Ratio		
	long port age	short port age	age diff	long port fee	short port fee	fee diff
Average	64.78	53.46	11.32*** (7.30)	0.49	0.60	-0.11*** (-7.20)

Table 9: Fama-MacBeth Analysis of ETF characteristics

This table reports estimates from Fama MacBeth 2-step regressions of ETF return on outperformance probability from previous global stock model, age of ETF measured in month and net expense ratio. The time-fixed effect is included.

ETF return	outperformance prob	t-stat	age	t-stat	fee	t-stat
	0.033*	(1.72)				
	0.033*	(1.81)	4.20×10^{-6}	(0.48)		
	0.034*	(1.72)			8.47×10^{-4}	(-0.39)
	0.035*	(1.81)	3.71×10^{-6}	(0.38)	7.20×10^{-4}	(-0.33)

Table 10: Other Model Settings with Validation

This table reports a robustness check for our global stock-trained model. We include classification-based and regression-based models. The classification-based models are linear model (logit model), decision tree, gradient boosting decision tree, and XGBoost, and the regression-based models are linear model (ordinary least square), decision tree, gradient boosting regression tree, and XGBoost. We evaluate the model on ETF markets and show the long return, short return, and long-short return with statistical significance. The '-' symbol is used to indicate the absence of results for the Gradient Boosting Regression Tree (GBRT) model.

	Classification Based			Regression Based		
Model	long ret	short ret	l-s mean	long ret	short ret	l-s mean
Logistic Regression	0.86%	0.19%	0.67%*** (3.16)	0.53%	0.40%	0.13% (0.53)
Decision Tree	0.90%	0.26%	0.64%** (2.48)	0.76%	0.50%	0.26% (1.37)
GBDT/GBRT	0.74%	0.37%	0.37%** (2.28)	-	-	- -
XGBoost	0.96	0.13	0.83%*** (3.36)	0.62%	0.30%	0.32% (1.59)

Table 11: Bond ETF Mareket

This table reports a robustness check on the bond ETF market for our global stock-trained model. We include classification-based and regression-based models. The classification-based models are linear model (logit model), decision tree, gradient boosting decision tree, and XGBoost, and the regression-based models are linear model (ordinary least square), decision tree, gradient boosting regression tree, and XGBoost. We evaluate the model on bond ETF markets and show the long return, short return, and long-short return with statistical significance.

	Classification Based			Regression Based		
Model	long ret	short ret	l-s mean	long ret	short ret	l-s mean
Logistic Regression	0.08%	0.13%	-0.05%	0.44%	0.06%	0.38%
			(-0.38)			(1.16)
Decision Tree	0.37%	-0.03%	0.40%**	0.04%	0.02%	0.02%
			(2.28)			(0.26)
Random Forest	0.16%	0.28%	-0.12%	0.24%	0.06%	0.18%
			(-0.33)			(1.19)
GBDT/GBRT	0.08%	0.13%	-0.05%	0.17%	0.07%	0.10%
			(-0.38)			(1.22)
XGBoost	0.06%	0.03%	0.03%	0.18%	0.10%	0.08%
			(0.39)			(0.68)

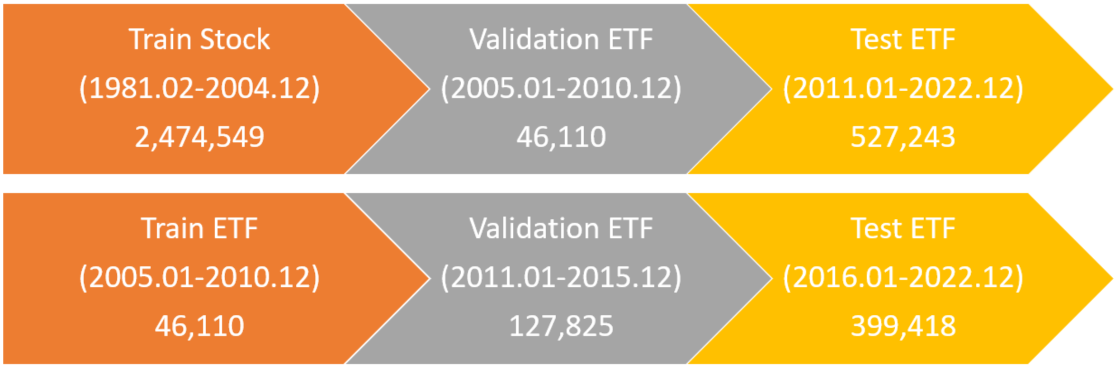

Figure 1: Train-Validation-Test Split

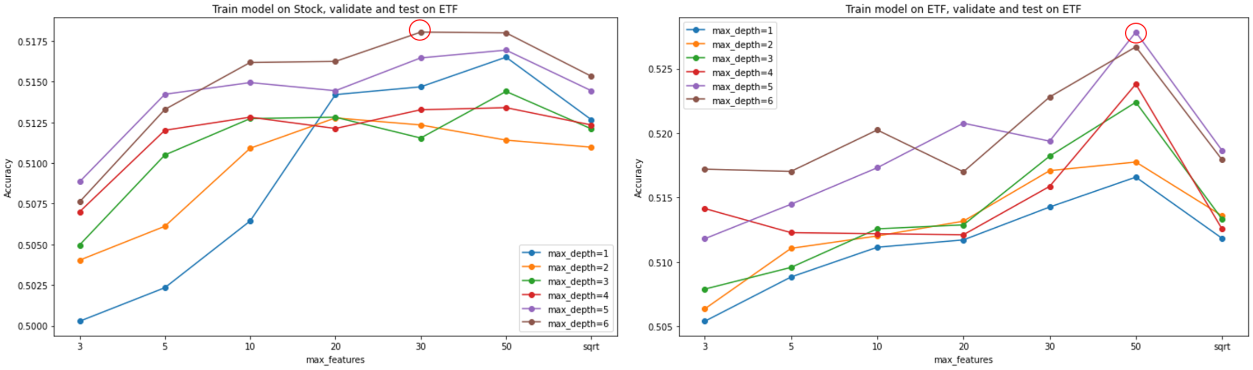

Figure 2: Hyperparameter Sensitivity Analysis

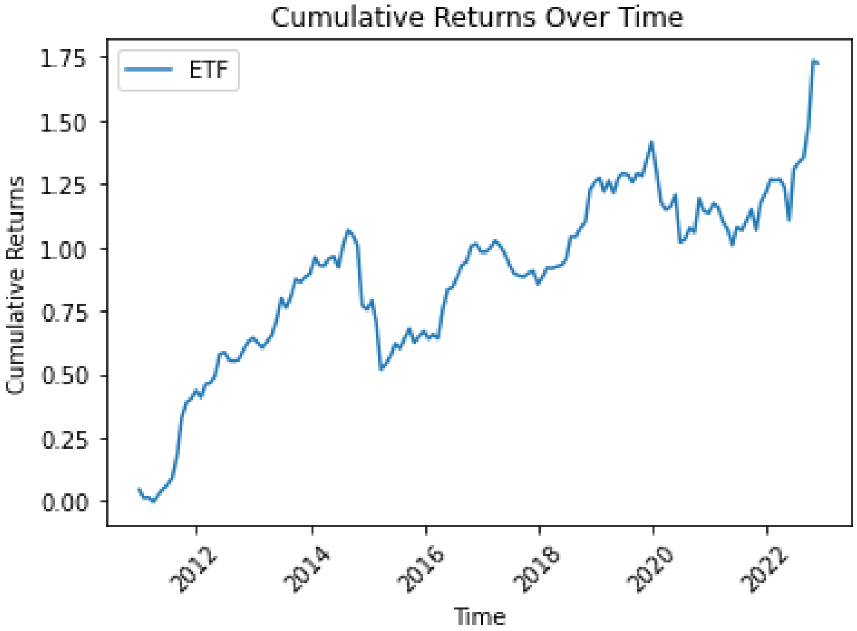

Figure 3: Cumulative ETF Returns Over Time

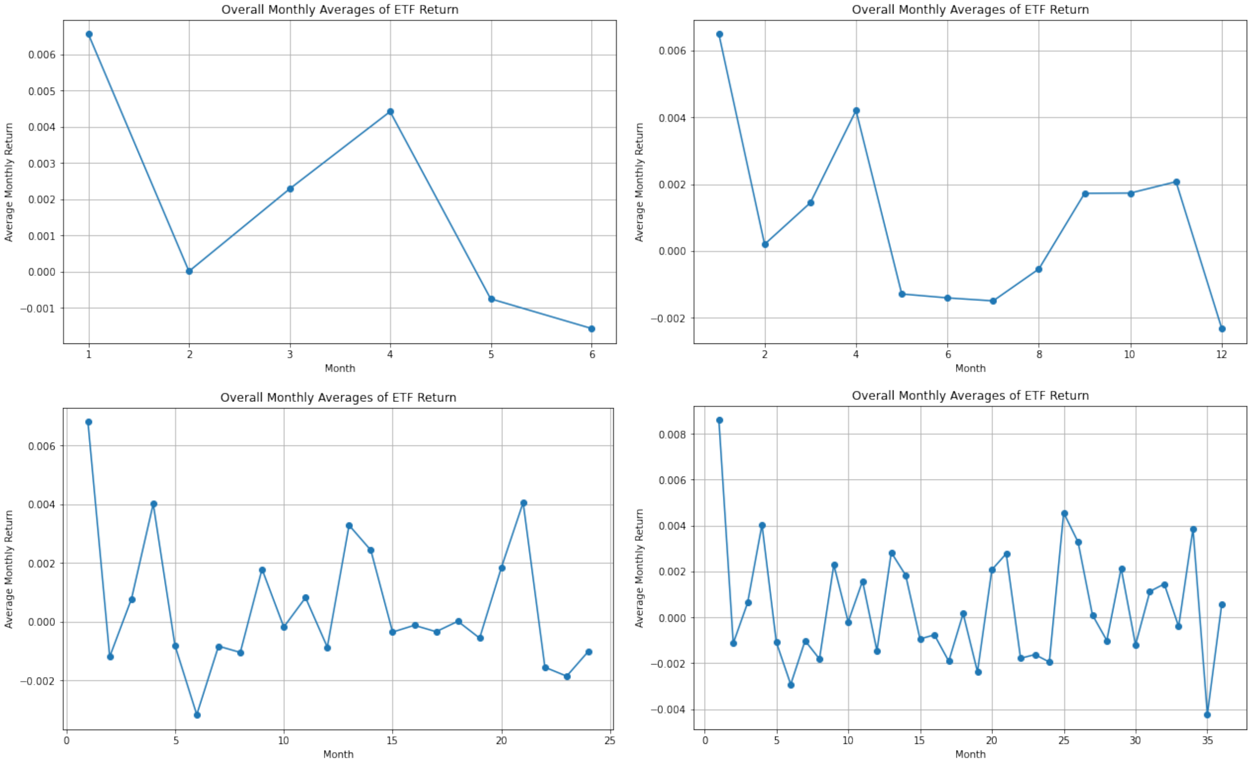

Figure 4: Different Rebalance Periods

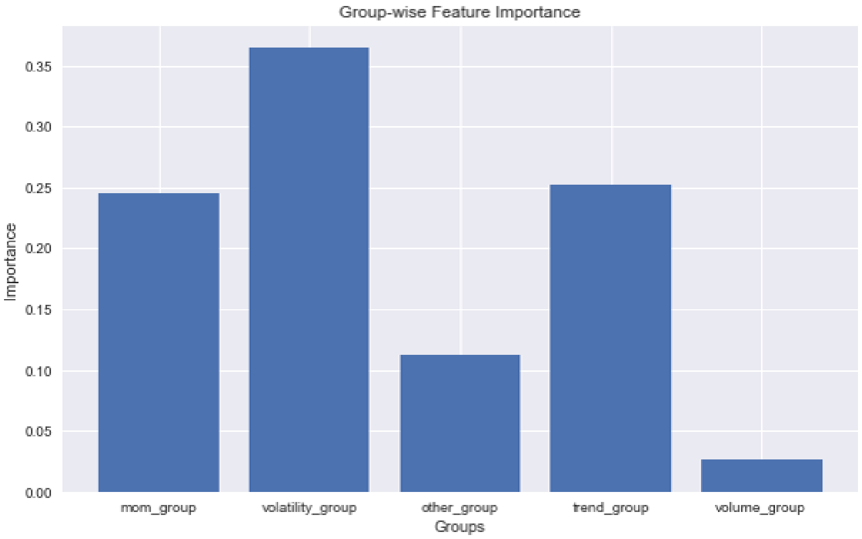

Figure 5: Grouped Feature Importance

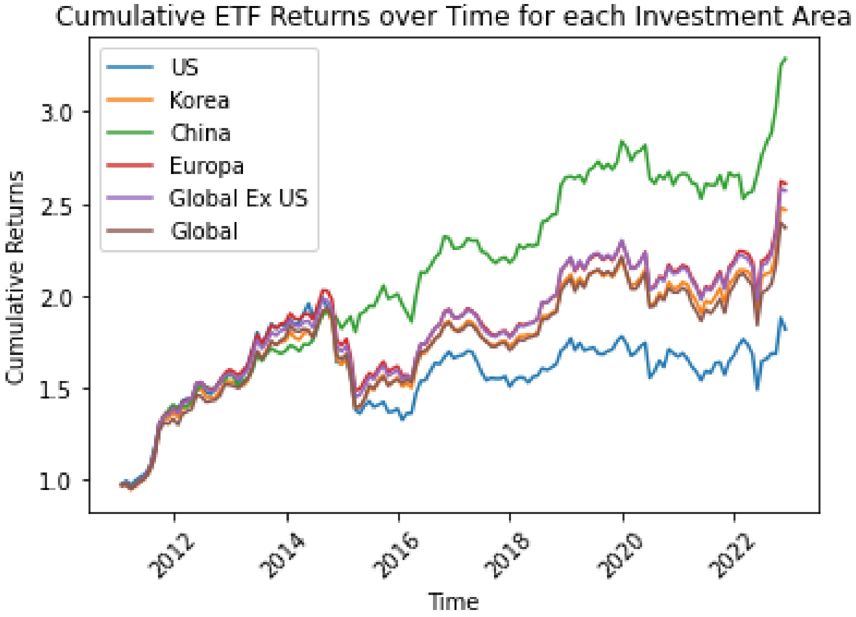

Figure 6: Cumulative ETF Returns Over Time In Different Areas

Appendix

Table A.1: Overview of Technical Indicators

Volume indicator	Volatility indicator	Trend indicator	Momentum indicator	Other indicator
volume_adi	volatility_bbm	trend_macd	momentum_rsi	others_dr
volume_obv	volatility_bbh	trend_macd_signal	momentum_stoch_rsi	others_dlr
volume_cmf	volatility_bbl	trend_macd_diff	momentum_stoch_rsi_k	others_cr
volume_fi	volatility_bbw	trend_sma_fast	momentum_stoch_rsi_d	others_dr_rel
volume_em	volatility_bbp	trend_sma_slow	momentum_tsi	others_dlr_rel
volume_sma_em	volatility_bbhi	trend_ema_fast	momentum_uo	others_cr_rel
volume_vpt	volatility_bbli	trend_ema_slow	momentum_stoch	
volume_vwap	volatility_kcc	trend_vortex_ind_pos	momentum_stoch_signal	
volume_mfi	volatility_kch	trend_vortex_ind_neg	momentum_wr	
volume_nvi	volatility_kcl	trend_vortex_ind_diff	momentum_ao	
volume_adi_rel	volatility_kcw	trend_trix	momentum_roc	
volume_obv_rel	volatility_kcp	trend_mass_index	momentum_ppo	
volume_cmf_rel	volatility_kchi	trend_dpo	momentum_ppo_signal	
volume_fi_rel	volatility_kcli	trend_kst	momentum_ppo_hist	
volume_em_rel	volatility_dcl	trend_kst_sig	momentum_pvo	
volume_sma_em_rel	volatility_dch	trend_kst_diff	momentum_pvo_signal	
volume_vpt_rel	volatility_dcm	trend_ichimoku_conv	momentum_pvo_hist	
volume_vwap_rel	volatility_dcw	trend_ichimoku_base	momentum_kama	
volume_mfi_rel	volatility_dcp	trend_ichimoku_a	momentum_rsi_rel	
volume_nvi_rel	volatility_atr	trend_ichimoku_b	momentum_stoch_rsi_rel	

Table A.1 Continued: Overview of Technical Indicators

volatility_ui	trend_stc	momentum_stoch_rsi_k_rel
volatility_bbm_rel	trend_adx	momentum_stoch_rsi_d_rel
volatility_bbh_rel	trend_adx_pos	momentum_tsi_rel
volatility_bbl_rel	trend_adx_neg	momentum_uo_rel
volatility_bbw_rel	trend_cci	momentum_stoch_rel
volatility_bbp_rel	trend_visual_ichimoku_a	momentum_stoch_signal_rel
volatility_bbhi_rel	trend_visual_ichimoku_b	momentum_wr_rel
volatility_bbli_rel	trend_aroon_up	momentum_ao_rel
volatility_kcc_rel	trend_aroon_down	momentum_roc_rel
volatility_kch_rel	trend_aroon_ind	momentum_ppo_rel
volatility_kcl_rel	trend_psar_up	momentum_ppo_signal_rel
volatility_kcw_rel	trend_psar_down	momentum_ppo_hist_rel
volatility_kcp_rel	trend_psar_up_indicator	momentum_pvo_rel
volatility_kchi_rel	trend_psar_down_indicator	momentum_pvo_signal_rel
volatility_kcli_rel	trend_macd_rel	momentum_pvo_hist_rel
volatility_dcl_rel	trend_macd_signal_rel	momentum_kama_rel
volatility_dch_rel	trend_macd_diff_rel	
volatility_dcm_rel	trend_sma_fast_rel	
volatility_dcw_rel	trend_sma_slow_rel	
volatility_dcp_rel	trend_ema_fast_rel	
volatility_atr_rel	trend_ema_slow_rel	
volatility_ui_rel	trend_vortex_ind_pos_rel	
	trend_vortex_ind_neg_rel	
	trend_vortex_ind_diff_rel	

Table A.1 Continued: Overview of Technical Indicators

trend_trix_rel
trend_mass_index_rel
trend_dpo_rel
trend_kst_rel
trend_kst_sig_rel
trend_kst_diff_rel
trend_ichimoku_conv_rel
trend_ichimoku_base_rel
trend_ichimoku_a_rel
trend_ichimoku_b_rel
trend_stc_rel
trend_adx_rel
trend_adx_pos_rel
trend_adx_neg_rel
trend_cci_rel
trend_visual_ichimoku_a_rel
trend_visual_ichimoku_b_rel
trend_aroon_up_rel
trend_aroon_down_rel
trend_aroon_ind_rel
trend_psar_up_rel
trend_psar_down_rel
trend_psar_up_indicator_rel
trend_psar_down_indicator_rel

Table A.1 Continued: Overview of Technical Indicators

20(12%)	42(24%)	68(40%)	36(21%)	6(3%)

Table A.2: Hyperparameters for All Methods

	number of estimators	max depth	max features
RandomForestClassifier	300	1,2,3,4,5,6	3,5,10,20,30,50,'sqrt'
DecisionTree	1	1,2,3,4,5,6	3,5,10,20,30,50,'sqrt'
GBDT/GBRT	300	1,2,3,4,5,6	3,5,10,20,30,50,'sqrt'
XGBoost	300	1,2,3,4,5,6	-

Table A.3: Overview of Investment Area

Investment Area	Number of ETFs
United States of America	2199
Global	1238
China	913
Europe	383
Korea	369
Global Ex US	296
Total	7639

Declaration

We acknowledge the use of ChatGPT [https://chat.openai.com/] to edit our writing at the final stage.

For part of the text, we entered the following prompts: “Improve my academic writing style.” “Is this paragraph grammatically correct?” and then we used the output to help revise our writing.