An Overview of Statistical Learning Methods for Factor-Based Country Allocation

Daniel Feng, Alex Starr, Shrimad Vora

Under the Oversight of Professor Tony S. Wirjanto

April 21, 2024

Table of Contents

1.1 Paper Overview

Our project encapsulates a unique area of factor investing that is less studied and reviewed than the traditional equity factor investing approach. By applying factor investing methods to asset allocation methods, this area of research aims to provide a basis for portfolio managers of more macro-focused investments to incorporate factor investing. We are looking primarily at global equity country allocation, however many of the same factor approaches apply to the fixed-income space as well as other asset classes. To continue research in this field, we plan to apply machine learning techniques to gain further insight into the significance of these research findings.

The primary goal of this research is to quantitatively determine if country-based equity returns exhibit the same dependence on the Fama-French three-factor model that traditional single-stock equity does. The paper, paper "Global Equity Country Allocation: An Application of Factor Investing" by Timotheos Angelidis and Nikolaos Tessaromatis shows this dependence qualitatively, by constructing portfolios consisting of of small-cap, value, high-momentum, and low-beta countries, and determining which factors will outperform the market. Our goal is to replicate these findings and apply a variety of statistical learning techniques as the basis for portfolio construction. We separately apply elastic net regression, decision tree methods, and a deep learning approach to pinpoint the relation between these factors and country-based return. This paper aims to provide a basis to assist asset managers in performing country-focused asset allocation by predicting outperforming countries against their world peers.

1.2 Literature Review:

The paper "Global Equity Country Allocation: An Application of Factor Investing" by Timotheos Angelidis and Nikolaos Tessaromatis extends the application of equity factor investing into the global equity country allocation space. The authors propose a factor model and strategy using country indices as the response variate, as opposed to the traditional individual equity as done in the Fama and French models. It further explores different portfolio construction and weighting techniques to constitute a “country index”. All in all, the paper aims to apply factor models to global asset allocation and to quantitatively provide a technique for determining country exposure. They demonstrate that a portfolio consisting of small-cap, value, high-momentum, and low-beta countries outperforms the total world market portfolio in terms of return and risk, with these factors being statistically significant. Additionally, the authors compare the performance of stock-based factor portfolios with country-based portfolios and, using multiple linear regression models, provide empirical evidence supporting the viability of country-based factor portfolios as an alternative implementation of factor investing. The paper also discusses the challenges of implementing academic research on factor portfolio construction, considering issues such as stock liquidity, transaction costs, turnover, and risk constraints faced by portfolio managers in practice. To further tie their results to the real world, the authors make use of empirical analysis using country ETFs as a stand-in for a country overall, and further, take into account ETF tracking error in their analyses. Overall, the study contributes to the understanding of factor investing and its application in global equity country allocation.

In the context of the factor investing research space, the paper references several academic works related to factor investing, including studies by Fama and French, Asness, Moskowitz, Pedersen, and others. It builds upon the factors identified by Fama and French and applies them to a new response domain. The authors compare the performance of country-based factors with the global factors of Fama and French and investable stock-based portfolios, providing insights into the potential benefits and challenges of country-based factor investing.

2.1 Data Preprocessing

As a representative for each country's equity market, we used the price level from its corresponding MSCI country index. We do so as each index is targeted to cover approximately 85% of its country's free float market capitalization, and is likewise a proxy for that equity market.

In following with the findings from the Angelidis and Tessaromatis paper we used a sample of 23 developed markets and 21 developing markets. The data is from April 2000 to the present as this was the inception date of many developing market MSCI indices. The only difference from the paper in the data collection is the exclusion of the MSCI Russia index which was discontinued in 2022, to prevent survivorship bias, we have included MSCI Russia until that date only. All indices have been converted to US dollar amounts to account for currency-induced fluctuations, this will make our findings more interpretable for single-currency asset managers.

All portfolios have been constructed based on the Fama-French three-factor model, we have likewise calculated each index factor as follows: value as the price to book value, size as the index underlying market capitalization, momentum as the trailing 12 months cumulative return, and a market risk premium based on the MSCI All World index as a proxy for "the market". All data was sourced from Bloomberg or calculated in methods in line with the Angelidis and Tessaromatis methods for factor calculation. Each MSCI index is reassessed monthly, and a forward log and simple return are calculated. For each input factor, uniformization was used to create a separate data field that can also be used as input.

3.1 Regression Methods

3.1.1 Elastic Net Regression

The goal is to build a model that can predict returns using Elastic Net regression, a technique that combines L1 and L2 penalties for variable selection and regularization. The Elastic Net model can be viewed as a linear combination of the Ridge and Lasso Models. The dataset we use contains factor data such as Size, Value, Momentum, and Market Risk Premium, along with corresponding returns and log returns.

After loading in and cleaning the dataset, rows with missing returns but complete factor data were identified, as these rows are suitable for prediction. After filtering the dataset, the inputs for our Elastic Net models were prepared by selecting relevant factors and target variables (1M_Simple_Ret and 1M_Log_Ret). A function that generates Elastic Net models was made to compare Elastic Net models with different alpha values, where alpha controls the regularization strength. We experiment with alpha values ranging from 0.01 to 1.0 to evaluate model performance.

We evaluated the generated Elastic Net models using mean squared error (MSE) and R-squared metrics. The models' performance is assessed across different alpha values. The goal was to highlight the trade-off between regularization strength and model fit, however, we concluded that any value chosen from (0-1) had a negligible impact on the MSE and R-squared error. To compensate for these findings, we perform a cross-validation grid search. The grid search explores combinations of alpha and l1_ratio parameters to find the best-performing model. The selected optimal model's parameters are then used for prediction. We found that a regularization strength of 1.0 and l1_ratio of 0.8 were optimal choices.

Using the chosen model parameters, we predict returns for new data points with missing returns but available factor data. The predicted returns and log returns are calculated, providing insights into potential country trends. The model predicts the Egypt Index (MXEG) to have the best return based on the chosen parameters for March 2024.

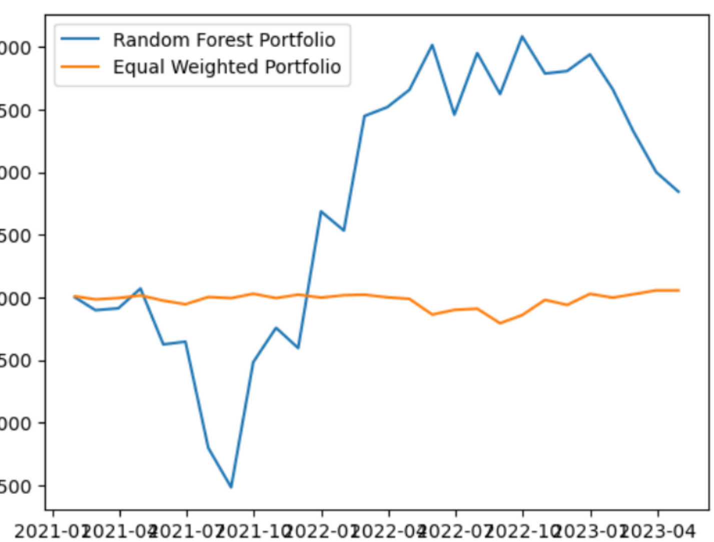
4.1 Random Forest Model

4.1.1 Forest Structure

The random forest model consists of an array of decision trees that branch off based on the potential amount of information gained. The amount of information gained is calculated as the difference between the entropy of the parent and the weighted average of the entropy of the left and right child nodes. Using the rudimentary random forest model coded, one uses it to forecast the log return of each period of each country index. Firstly, one separates the data into training data and test data using a specific date value and trains and builds the random forest model with 50 decision trees; each tree is trained from a sample randomly selected from the training data using bootstrapping. Afterwards, the 50 trees would decide the best possible answer using a majority voting scheme and output an appropriate return value for each country index at each timestep.

4.1.2 Model Implementation

After receiving the predicted returns, one conducts an MSE analysis to gauge the model's efficiency. The average MSE for predicting the returns of 43 (exclude MXRU since the index stopped updating in 2022) country indices is around 0.0093, which is a relatively decent accuracy. To gain a better insight into the model's effectiveness and the predicted returns, one constructed a global portfolio and compared the profit gained against an equally weighted global portfolio. Both portfolios contain the 43 indices at first. The equally weighted global portfolio would hold onto all indices and dynamically rebalance to ensure all indices have the same weight. On the other hand, the new global portfolio would sell the 14 indices with the lowest predicted returns and use the money to buy the 14 indices with the highest forecasted returns. One uses the data starting in 2021 to avoid potential noises caused by the Pandemic, and one realizes that the new portfolio significantly outperforms the equally weighted portfolio more than half the time.



Additionally, the random forest model built runs with a max depth of 100, a min split of 2, and 50 trees in the forest. While the model assumes that these parameters give the best results when predicting all country's indices, it's important to note that this may not always be the case. However, there is room for improvement. By performing hyperparameter tuning, one can produce the ideal set of parameters to forecast each country index, giving more accurate predictions. Not to mention, there is a world of options for achieving even better results. Additional parameters and other processes, such as pre and post-pruning, can further enhance the model's performance.

5.1 Deep Learning

5.1.1 Model Architecture

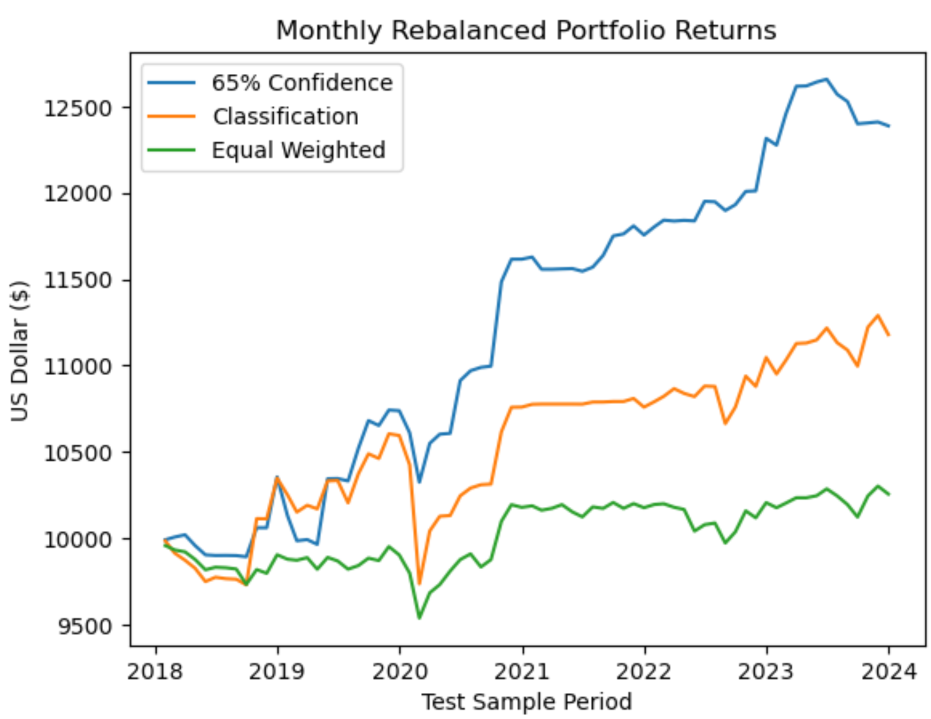
Two neural networks have been used to form predictions, that vary slightly in implementation. Each network is a dense neural network containing three dense layers with the first two using ReLU activation. Between each layer is a 10% dropout, which was chosen to prevent overfitting in the training sample. To take into account the country, a categorical variable is added with the country in a one-hot vector form, this is included in the model.

The first model is set to predict a three-class return profile, each corresponding to 33% quantile of the training set return. This classification breakdown was chosen to potentially provide a buy, sell or hold signal. The final layer uses a softmax activation layer to map the final three-class vector to a probability vector, and is trained with categorical cross-entropy loss. The category with the highest probability is selected for that signal.

The second model follows the same architecture but instead aims to predict market outperformance. The 3-month log return variable is mapped to a classification of size two, whether each return outperforms the equal-weighted portfolio or not. Since this is now a two-classification the model uses a sigmoid activation layer with binary cross-entropy loss to train. The predicted value is thus the probability that the forward return is greater than the market average return.

5.1.2 Model Assessment

To assess the effectiveness of each model, we created monthly rebalancing portfolios consisting of MSCI indices. The control portfolio consists of each MSCI index given equal weight and rebalanced to that weight monthly. The two other portfolios are also rebalanced monthly, but only countries with predicted positive returns are put into the index. For the first model, this means having a BUY signal for that index, for the second model, this means having a predicted probability of positive return of at least 0.65. The threshold of 0.65 was chosen to only select those that the model is most confident about. In effect, every month we selected countries (likewise perform an asset allocation decision) that we think will outperform the market.



The portfolio rebalancing essentially "chooses" or predicts which countries it believes will perform well over the next period. The above chart indicates the cumulative return of $10,000 invested in each portfolio strategy. Of note the two deep learning models (in orange and blue) outperform the equal-weighted portfolio (in green). To avoid fallacious cumulative returns, the above backtest assumes $10,000 paid allocated at every rebalancing period, with profit or loss added externally. All together, the above models adequately used factors from the three factors model to provide an edge in outperforming the market and making country-based asset allocation decisions.

6.0 Findings

The goal of this project is to provide a statistical learning approach to country-based factor investing. By gaining an understanding of the data from elastic net regression, and then a foray into interpretability with decision trees, we have arrived at a neural network model that provides a "leg-up" in predicting future returns on a country-level scale. The work with decision trees and random forests aims to break down the returns of each country and to provide an alpha by separately reconfiguring the portfolio. The neural network approach predicts each return with its country as a categorical input. Going forward, combining these different models into a single model may more adequately be able to predict forward returns. By providing a framework for country-based asset allocation, we find that asset managers can in fact apply factor investing, such as the Fama and French three-factor model, as a way to gain excess return in the market.

7.0 Appendix

Neural Network Model: Country Allocation Neural Network.ipynb

Decision Tree Model: RandomForest.ipynb

Multiple Linear Regression Models: mlr.ipynb

Data Source: CountryFactorData.csv