AFM 423 – Topics in Financial Econometrics
Group Project #1

Group Member:

Samuel Benedict (sbenedict@uwaterloo.ca)

Proposed Topic:

Application of Supervised Machine Learning Models to Improve Equity Return Prediction and Portfolio Construction in Factor Investing

Introduction

Since the discovery of the three-factor model by Fama and French in 1993, factor investing has become a prominent trend in asset management. However, new financial research papers have proven that conducting linear regression on these three factors is insufficient to explain asset returns, hence the need to incorporate additional potential factors that can capture more variability in asset returns (Fama & French, 2015). As discussed in the literature review section below, these additional factors will present multicollinearity, where the factors might be correlated to each other and would violate the assumptions of a linear regression model. Moreover, these factors might not impact the returns linearly as assumed by linear regression.

With new advancements in statistics, novel machine learning models can now be used as a method to predict asset returns. For example, regularization can perform feature selection which minimizes the collinearity between various factors. Neural networks and tree-based methods can use the factors non-linearly to predict stock returns. Specifically, recurrent neural networks can be used to give a better prediction on sequential data. Finally, tree-based methods can display the feature importance, giving insight into important factors in the prediction.

Additionally, it would be interesting to see how these factor models can be used in constructing portfolios that would bring excess return compared to the market. Using the portfolio sorting as conducted by Fama and French (1993), we can test whether the machine learning predictions can be used to construct a profitable long-only or long-short portfolio strategy.

Project Goal

This project has two main implementation goals:

1. Check if the implementation of machine learning (Regularization, Random Forest, XGBoost, Neural Networks, or Recurrent Neural Networks) can improve the prediction of equity returns compared to using standard multiple linear regression.

2. After obtaining the return prediction and sorting the stocks into multiple portfolios, check which of the portfolio construction strategies (Long-Only or Long-Short) can result in the highest portfolio return.

Literature Review (Papers on Factor Investing)

"Common Risk Factors in the Returns on Stocks and Bonds" by Eugene F. Fama and Kenneth R. French (Journal of Financial Economics, 1993, p.3-56)

As one of the earliest papers to challenge the CAPM and include factors other than market risk premia in explaining asset returns, this paper is a cornerstone of modern factor investing. The simple and intuitive factor model also becomes one of the few reasons why this paper is a foundational reference and a benchmark to beat for future research in factor investing. This paper is built upon Fama and French's own research in 1992, which states that market return does not fully explain the average stock returns. Instead, by adding size (market capitalization) and book-to-market value (BE/ME), then it explains the average stock returns better, as confirmed by testing it across stocks listed on NYSE, Amex, and NASDAQ from 1963 to 1990 (Fama & French, 1993, p.4). This argument can also be extended to accommodate the bond market using 2 additional factors: term and default risk. In the case of stocks, it was concluded that adding term and default risk does not significantly improve the prediction made by just using the 3 factors (Fama & French, 1993, p.41-42).

One important distinction between this paper and its predecessor is that it uses a time-series regression (Black-Jensen-Scholes approach) instead of a Fama-Macbeth regression. Monthly returns of stocks are regressed on returns from a portfolio of stocks that are the top or bottom quintile in size or book-to-market ratio. This top/bottom return difference is later referred to as SMB (Small – Big) for size and HML (High – Low) for BE/ME. Using this regression alongside the three factors, the paper confirms that this model captured most of the cross-section stock returns (Fama & French, 1993, p.51). Moreover, the residuals from this three-factor regression performed well not just in capturing the risk factors (market, size, and BE/ME), but also in isolating the "firm-wide components of returns" (Fama & French, 1993, p.51-54).

Another strength displayed by this regression is that it is statistically robust. This is proven by checking if the basic regression assumptions hold and confirming that all 3 factors are necessary to explain the variation of stock returns (Fama & French, 1993, p.41). Additionally, size and BE/ME are nearly uncorrelated, and these risk factors are only weakly correlated to the market risk factor (Fama & French, 1993, p.26). However, a major drawback of this paper is that it implicitly assumes these risk factors to build upon each other linearly by using the linear regression model, which most likely is not the case practically. In addition, the paper mentioned that the size and BE/ME factors may not be able to address the impact of earnings or profitability on stock returns (Fama & French, 1993, p.55).

"A Five-Factor Asset Pricing Model" by Eugene F. Fama and Kenneth R. French (Journal of Financial Economics, 2015, p.1-22)

Building on the previous 3-factor model, new research by Novy-Marx (2013) and Titman (2014) found that there are still anomaly variables, where the model does not explain some portion of

the variance of average stock returns. In response, Fama and French now augment the model to accommodate 2 additional factors, which are profitability and investment. Profitability specifically refers to the robustness of the company's profitability, while investment refers to the aggressiveness of the company's asset expansion. Continuing the spirit of their prior research, the authors take the difference between returns from the top and bottom quintiles for both profitability RMW: Robust – Weak) and investment (CMA: conservative – aggressive) before adding them alongside the 3-factor regression model (Fama & French, 2015, p.3).

The paper begins by creating different sorts of portfolios with various factor combinations and conducting regression on the different sorts. A high correlation was found between HML, RMW, and CMA on various sorts, with their correlation in 2x2 sorts being north of 0.95 (Fama & French, 2015, p.8). This suggests that there is multicollinearity between these 3 factors that must be eliminated to prevent the regression X matrix from being near singular. The authors also found that HML is a redundant factor when they consider RMW and CMA. Moreover, adding HML would not improve the efficient portfolio made from other factors (Fama & French, 2015, p.12). Hence, the authors suggest either dropping HML or replacing it with HMLO, which is an orthogonal HML that is obtained by regressing HML with excess market return (Fama & French, 2015, p.12).

The additional factors allow the model to explain between 71 to 94% of the cross-section return variance, a considerable improvement from the 3-factor model (Fama & French, 2015, p.17-19). This model is also backed by robust model diagnostics and summary statistics. However, some anomalies are left unexplained, particularly on the returns of small-cap stocks. For example, stocks with negative RMW exposure are expected to gain fewer profits, but small-cap stocks with RMW exposure do not necessarily have low profitability. As summarized at the end of the paper, "the most serious problems of asset pricing models are in small stocks" (Fama & French, 2015, p.19).

"Value and Momentum Everywhere" by Clifford S. Asness, Tobias J. Moskowitz, and Lasse H. Pedersen (Journal of Finance, 2013, p.925-984)

In this paper, the authors argued that both long-run value and recent performance (momentum) play a crucial role as significant return factors across various assets globally. Specifically, value is specified as BE/ME, while momentum is described as the cumulative raw return in the past 12 months (Asness et al, 2013, p.936). The significance of both factors is backed by consistent evidence that value and momentum generate return premiums across equities, government bonds, currencies, and commodities (Asness et al, 2013, p.930).

This paper proposes a 3-factor model consisting of market, value, and momentum, with a similar regression method as implemented in Fama & French's 1993 paper. The authors found that value and momentum are negatively correlated, and that momentum strategies are correlated globally. Moreover, this model can capture 71% of the variation in returns across 48 assets, signifying the viability of this model to predict asset returns. However, when tested against 5x5 sorts of size/value and size/momentum, it is better to use the original three factors (market, size, value) plus momentum (Asness et al, 2013, p.966-969). This implies that there is still room to add value and momentum on top of existing factors.

Another downside presented by this paper is that it only uses the simplest value and momentum metrics, which most likely does not reveal the true impact of these factors (Asness et al, 2013,

p.978). On the other hand, if assets with the highest predicted returns are grouped into a long-only portfolio, this would still provide "abnormal returns", signifying the edge this model has compared to the market portfolio (Asness et al, 2013, p.977). This return would be magnified by simultaneously shorting assets with the lowest predicted returns, although this does not guarantee superior returns in practice due to the unaccounted presence of transaction and shorting costs (Asness et al, p.975-976). Finally, the authors suggest combining both value and momentum factors in a model instead of analyzing it separately, as it would yield a higher portfolio return and Sharpe Ratio (Asness et al, p.945).

"Betting Against Beta" by Andrea Frazzini and Lasse H. Pedersen (Journal of Financial Economics, 2014, p.1-25)

This paper starts off by observing the growing presence of portfolio tilts toward high-beta assets, which suggests that high-beta assets would require lower risk-adjusted returns compared to low-beta ones (Frazzini & Pedersen, 2014, p.2). Hence, the authors introduced the betting-against-beta (BAB) factor, which consists of leveraging stocks with low beta until the beta is 1 and shorting stocks with high beta until the net portfolio beta is 0. In most cases, this factor portfolio would be able to generate higher alpha and Sharpe Ratio compared to traditional market portfolios (Frazzini & Pedersen, 2014, p.11-13). Additionally, across different asset classes tested, the models that are equipped with the BAB factor managed to gain an edge on their returns (Frazzini & Pedersen, p.15).

Having said that, there are 2 significant drawbacks to this paper. First, the factor relies on a relatively efficient market with no shorting restrictions. When there are liquidity constraints, the BAB factor would realize negative returns, as the beta across various assets is compressed towards 1, and that shorting would not add any edge to portfolio returns (Frazzini & Pedersen, 2014, p.21). This is a phenomenon commonly referred to as beta compression. Lastly, the paper conducts no checks of multicollinearity when trying to combine BAB with other factors in a predictive model. This may cause a weaker prediction since BAB might be closely correlated with other factors, which would cause the model to implode.

General Methodology

Based on our literature, the factors that we will be considering are:

1. Market (Rm – Rf: Market – Risk-Free Return)
2. Size or Market Capitalization (SMB: Small – Big Market Cap)
3. Book-to-Market Value (HML: High – Low BtM)
4. Momentum (WML: Winners – Losers)
5. Profitability (RMW: Robust – Weak Profits)
6. Capital Investment (CMA: Conservative – Aggressive Investments)
7. Betting Against Beta (BAB: Long Low Beta, Short High Beta)

Meanwhile, the predictors that we will be considering are:

1. Linear Regression
2. Linear Regression with Regularization (Elastic Net)
3. Random Forest

4. XG Boost
5. Neural Network
6. Recurrent Neural Network (ex. LSTM)

In general, this project will be executed in the following steps:

1. From the data_ml dataset, keep the monthly return column (R1M_USD) and columns whose features are used to construct our factors (ex. Mkt_Cap_3M_USD, Bv, etc.).
2. For each factor (except the market factor), sort the stocks by the factor value. Then, group them into 5 quintiles.
3. Take the average monthly returns from the top and bottom quintiles. Then, take the differences as applicable.
4. Split the portfolio into sliding train/validation and testing sets. For example, if we have N months, we can use months 1 to k ($k < N$) to predict returns on month k+1. Next, use months 2 to k+1 to predict month k+2. We do this until we can predict month N.
5. On each training step, train and validate the data using different predictors. Then, we compare the predicted return to the actual return for each month. Aggregate the results for each predictor to get an understanding of the predictive accuracy.
6. For each predictor, sort the stocks by average predicted return and group them into 5 quintiles. Then, apply a long-only strategy on the top quintile, as well as a long top quintile / short bottom quintile strategy.
7. Compare the risk and return of the strategies to get an understanding of the portfolio profitability.

Model Assessment Metrics

We can use the following metrics to assess if our model is better than the benchmark (standard linear regression model with all the factors):

1. Correlation Matrix between factors and F-test for regression models
2. Root Mean-Squared Error (RMSE) of predicted returns vs actual return
3. Mean Absolute Error (MAE) of predicted return vs actual return

Meanwhile, we can use the following metrics to determine the risk and reward features of long-only and long-short portfolios from the prediction:

1. Portfolio return
2. Portfolio volatility (standard deviation)
3. Portfolio Sharpe Ratio

References

Asness, C. S., Moskowitz, T. J., & Pedersen, L. H. (2013). Value and Momentum Everywhere. *Journal of Financial Economics*, 925-984.

Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns of Stocks and Bonds. *Journal of Financial Economics*, 3-56.

Fama, E. F., & French, K. R. (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, 1-22.

Frazzini, A., & Pedersen, L. H. (2014). Betting Against Beta. *Journal of Financial Economics*, 1-25.

The grade is allocated to GPR1 according to these criteria:

---

1. Suitability of Topic: Very good
2. Suitability and Quality of References: Very good
3. Cited Weaknesses and Strengths of each References: Very good
4. Synthesis of Weaknesses and Strengths across References: Good
5. Proposed Remedies in terms of Scope/Breadth: Very good
6. Proposed Remedies in terms of Depth: Very good
7. Proposed Remedies in terms of Doability: very godd
8. Proposed Remedies in terms of Novelty: Very good

---

9. Final Comment: This is an outstanding report
10. Grade: **96/100**
