Application of Machine Learning in Fama-French Five-Factor Model

Melody Mao, Simon Zhang

AFM 423 | Winter 2024

Instructor: Tony S. Wirjanto

March 3rd, 2024

Word Count: 1160 words

1. Introduction

With the development of financial economics, the analysis and prediction of stock market returns have always been among the most captivating topics in both academia and the industry. Sharpe (1964) and Lintner (1965) proposed the Capital Asset Pricing Model (CAPM), which determined the relationship between return and risk, laying the foundation for asset pricing models. However, CAPM's single-factor approach often fell short of fully explaining the complexities of stock returns in practical applications.

Seeking to enhance CAPM's accuracy and overcome its limitations, Fama and French (1993) introduced two new factors—size and value—and proposed the Fama-French Three-Factor Model. This model marked a significant advancement in asset pricing theories. Further addressing anomalies, Fama and French (2015) expanded their model to include two additional factors, profitability and investment pattern, culminating in the Fama-French Five-Factor Model.

However, the explanatory power of the Fama-French Five-Factor Model remains controversial. In this report, we will discuss application of machine learning, especially Random Forest and Support Vector Regression, within the Fama-French Five-Factor Model, comparing their explanatory effects to those of the traditional multiple linear regression model.

2. Literature Survey

2.1 Fama-French Five-Factor Model: *A five-factor asset pricing model*

The paper extends Fama-French Model from three factors to five factors by adding profitability and investment factor to enhance explanatory power for variations in average returns related to

profitability and investment. RMW is the difference between robust and weak profitability, while CMA is the difference between conservative and aggressive investment pattern. According to the test, Five-Factor Model significantly outperforms the Three-Factor Model, explaining between 71% and 94% of the cross-section variance of expected returns.

The paper suggests that HML might be a redundant factor when considering RMW and CMA. T-statistics in HML regression for different portfolios are all fall to reject. Furthermore, the research indicates the main problem of Five-Factor Model is that it fails to fully explain the lower average returns on small stocks with large investment but low profitability.

The strength of the paper is its comprehensive testing across various factor combinations - 2x2 factors, 2x3 factors, and 2x2x2x factors – rather than solely examining the five factors together. This comprehensive testing allows us to discover the redundant factor and specific types of stocks that the model fails to explain, providing deeper insights into the Five-Factor Model.

However, this research restricts the model implement in classical multiple linear regression. With the maturity of Machine Learning, we could implement them in the Five-Factor to improve improve beta estimation and delve into HML factor and small stocks with large investment but low profitability.

2.2 Random Forest: *A Machine Learning Approach to Risk Factors: A Case Study Using the Fama–French–Carhart Model*

The paper first discussed the limitation of linear and parametric non-linear regression models used in the financial world, nowadays. According to this paper, for a linear regression model, it restricts itself to a small number of factors and it may provide an inadequate picture of portfolio behavior over a given measurement period. Moreover, these expanded models do not offer maximum informativeness as they fail to adequately represent nonlinear behaviors and interaction effects among the factors. This limitation highlights the inability of extended linear models to fully encapsulate and accurately reflect the complexities inherent in financial data behaviors and relationships. For parametric nonlinear models, it also has some shortcomings such as dependence on sample data, analytical derivation and convergence issues (Simonian, Wu, Itano, & Narayanam, 2019).

In order to deliver a solution that could avoid these limitations and shortcomings, the author introduced Random Forest (RF) algorithms. The authors demonstrate the utilization of the RF to create models within a unified framework that elucidates the sensitivity of assets to various factors, similar to those identified by traditional models but with enhanced explanatory capabilities. RF-based models adeptly capture nonlinear relationships, abrupt changes (like threshold correlations), and variable interactions without the complexity of intricate functional forms or extra interaction terms, adhering to the principle of parsimony. This approach allows for a more nuanced and potent analysis while maintaining simplicity and efficiency in model construction.

Although there are many benefits plotted above, there are some drawbacks that need to be considered while applying RF to Fama-French Five factors. One is that we only have five factors

as a base model. Therefore, after applying RF, the model might still underfit. Second is it is becoming hard to interpret the result. Unlike linear regression models, the relationship between factors and dependent variables is really straightforward. However, when it comes to RF, it is not apparent and it is hard for decision makers to trust the result.

2.3 Support Vector Regression: *A MACHINE LEARNING APPROACH TO THE FAMA-FRENCH THREE- AND FIVE-FACTOR MODELS*

Support Vector Regression (SVR) was introduced by Cortes and Vapnik in the year of 1995. It is a supervised learning approach and is considered as a strong statistical tool that is used for model prediction. First, SVR employs the concept of structural risk minimization, which helps to reduce the model's prediction error bound, enhancing its performance, particularly in situations with limited data points and features. Second, SVR has the flexibility to model both linear and non-linear relationships between the dependent and independent variables, allowing it to adapt to various data structures and uncover complex patterns within the data.

According to this research paper (Diallo, Bagudu, & Zhang, 2019), “Fama-French BSVR three-factor estimations attain out-of-sample (testing dataset) cor- relation coefficients of 94% for portfolio returns for the consumption and manufacturing industries. A correlation of 92% between the predicted and experimental values of portfolio returns was found for the high-tech industry; 91% was found for the mining, construction, transportation, hotels, entertainment, and finance industries.”

One thing concerned is some parameters (C, and) are required to tuning the model. Therefore, Bayesian algorithms might be required to find these parameters.

3. Method

3.1 Data

Stock return data will be collected from data_ml.xlsx in AFM 423 class folder, from November 1998 to March 2019. Factor variables will be collected from Kenneth French Data Library, including:

- Rm-Rf: Market return – risk free rate
- SMB (Small Market Cap – Big Market Cap): Size factor
- HML (High PB – Low PB): Value factor
- RMW (Robust profit – Weak profit): Profit factor
- CMA (Conservative – Aggressive): Investment Factor

The Fama-French Five-Factor Model has an equation:

$$R_{I,t} - rf_t = \alpha_i + \beta_{i,MKT} MKT_t + \beta_{i,SMB} SMB_t + \beta_{i,HML} HML_t + \beta_{i,RMW} RMW_t + \beta_{i,CMA} CMA_t$$

3.2 Processing

1. Data will be divided two parts:
    a. Part I (training sample), from November 1998 to January 2014, to determine parameters of each model.
    b. Part II (testing sample), from February 2014 to March 2019, to evaluate prediction error.

2. Implement classical multiple linear regression model, Random Forest, and Suppose Vector Regression, given R packages
3. Prediction error will be estimated by:
    a. RMSE: Root Mean-Squared Error
    b. MAE: Mean Absolute Error
4. Compare and Visulize explanatory effects of classical multiple linear regression model, Random forest, and Support Vector Regression

Reference

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. https://doi.org/10.1016/j.jfineco.2014.10.010

Simonian, J., Wu, C., Itano, D., & Vyshaal Narayanam. (2019). A Machine Learning Approach to Risk Factors: *A Case Study Using the Fama–French–Carhart Model*. *The Journal of Financial Data Science*, *1*(1), 32–44. https://doi.org/10.3905/jfds.2019.1.032

Diallo, B., Aliyu Bagudu, & Zhang, Q. (2019). A Machine Learning Approach to the Fama-French Three- and Five-Factor Models. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3440840