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

Melody Mao, Simon Zhang

School of Accounting and Finance, University of Waterloo, Canada

Abstract

This report explores the application of machine learning models, especially Random Forest (RF) and Support Vector Regression (SVR), within Fama-French Five-Factor (FF5) Model. Traditionally, the FF5 model was implemented in multiple linear regression. However, with the increasing complexity of the financial market, machine learning models have shown promise for enhancing predictive accuracy. We compare RF and SVR against the linear regression by training on the past 65 months of historical data and testing stock returns from February 2014 to March 2019. Our results indicate that RF and SVR outperform linear regression with lower MSE, lower MAE, and higher Hit Ratio. These findings underscore the potential application of machine learning models in the FF5 model.

1. Introduction

Before the FF5 model, the Fama-French Three-Factor (FF3) model was developed to revise the traditional Capital Asset Pricing Model. It captures the relationship between excess returns and a combination of market risk, company size, and book-to-market value. Later, in 2015, Fama and French extended this model to include two additional factors: profitability and investment patterns. Although the FF5 model outperformed the FF3 at that time, the limitations of the linear regression model were still apparent. Unlike linear regression models, machine learning algorithms can capture the non-linear relationships between factors and excess returns. Additionally, the Random Forest model is less likely to be influenced by correlations between factors. Therefore, the problems existing in Ordinary Least Squares (OLS) regression models are unlikely to occur in Random Forest models. Furthermore, the third model used in this report, Support Vector Regression (SVR), is another robust prediction model. It not only inherits the benefits of the Random Forest but also offers an excellent balance between overfitting and underfitting. In this report, both Random Forest and SVR are applied to FF5 factors to replace the linear regression model, which is generally used on these factors to capture more complex relationships between factors and excess return on assets.

2. Variable and Measures

In this research project, considering FF5 factors, we select the following predictors and define our response variables correspondingly.

- Predictors
    - Rm-Rf (Market return - risk-free rate): Market returns
    - 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

These factors are downloaded from the Kenneth French Data Library and serve as our independent variables.

- Response Variables
  - Monthly return: predict portfolio return in the next month
  - R1M_Usd: returns forward 1 month

  This monthly return is predicted by FF5 factors and their betas based on linear regression or training results, and the R1M_Usd is extracted from data_ml.RData.
- Measurement of Response Variables
  - Mean-squared error (MSE): measures the average squared difference between predicted return and the actual return.

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

  - Mean Absolute Error (MAE): measures the average magnitude of the errors in a set of predictions.

$$\frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

  - Hit Ratio: measures the accuracy of the model's predictions in terms of directionality.

  We implemented the above metrics to assess the fitting performance of each model (Multiple Linear Regression, RF, and SVR).

3. The Application of the ML approach to Factor Investing

In this project, we focus on the application of two specific machine learning models, Random Forest (RF) and Support Vector Regression (SVR).

3. 1 Random Forest

Random Forest is an ensemble regression method that constructs multiple decision trees by randomly selecting subsets of features and data points through bootstrapping. This method helps in reducing model variance and preventing overfitting. The predicted outcome is derived from the average output across all trees:

$$\widehat{y_i} = \frac{1}{n}\sum_{i=1}^{n} tree_i(x)$$

Additionally, Random Forest provides an output known as feature importance (FI), which assesses the contribution of each factor (size, value, profit, and investment) to the predicted stock returns.

3.2 Support Vector Regression

Support Vector Regression (SVR) extends the concepts of Support Vector Machines (SVM) from classification to regression tasks. Unlike traditional regression models that aim to minimize the error between the predicted and actual values, SVR seeks to fit the prediction error within a predefined threshold. This is achieved using a concept known as the epsilon-insensitive tube, which essentially ignores errors within a certain range, allowing the model to focus on more significant trends rather than minor deviations.

4. The experimental methodology

The project begins by fitting historical data from the dataset into a linear regression model, which is then used to predict future excess returns. Subsequently, the report utilizes the same dataset but applies both Random Forest and Support Vector Regression to the FF5 factors instead of a linear model. At the end of the process, the report compares the predictive performance of these three models.

4.1 Data Size

In this project, we utilize data from the past 65 months (about 5 and a half years) for training purposes and testing returns from February 2014 to March 2019. It is a trade-off between execution time and acquiring enough data points. The choice of 65 months (about 5 and a half years) is based on historical statistics that the average economic cycle in the U.S. lasts approximately five and a half years. This testing period provides a comprehensive view across various economic stages. Additionally, the execution time for processing a 65-month period is feasible in our testing.

4.2 Fama-French Five Factor Model

Several data sources need to be highlighted. First, the FF5 factor model is described by the equation,

$$Ri_t - RF_t = \alpha_i + \beta_{MKT}(Rm_t - RF_t) + \beta_{SMB}SMB_t + \beta_{HML}HML_t + \beta_{RMW}RMW_t + \beta_{CMA}CMA_t + ei_t$$

The right-hand side of this equation can be calculated from $R1M\_Usd - RF$, where $R1M\_Usd$ is return froward 1 month which can be retrieved from the data_ml dataset, and RF can be obtained from the Ken French Data Library. Rm, RF, SMB, HML, RMW, and CMA can all be retrieved

from the Ken French Data Library. For each stock at time t, we use historical data from the period [t-65, t] to build a linear model and predict the stock price at time t for that stock. In the end, the dataset comprises predicted values and actual excess returns.

4.3 Random Forest

The Random Forest algorithm was applied similarly to the traditional multiple linear regression model. The main issue is determining the number of trees in the model. A random forest with too many trees can lead to excessive computation times without a corresponding performance improvement, while too few trees may not capture relationships accurately. Based on our testing, we have set the number of trees to 15.

4.4 Support Vector Regression

Support Vector Regression was implemented similarly to the traditional multiple linear model as well. The trControl parameter in SVR specifies various training control options for the model training process. It defines how the model should be trained, including how resampling should be handled to estimate model accuracy. To enhance the robustness of this model and ensure it generalizes well on new, unseen data, a 4-fold cross-validation scheme is applied. This method divides the data into four equal parts, using each in turn for validation while training on the remaining three. Such a strategy helps in mitigating overfitting and provides a more accurate estimate of the model's performance on independent data set.

5. Results and discussion

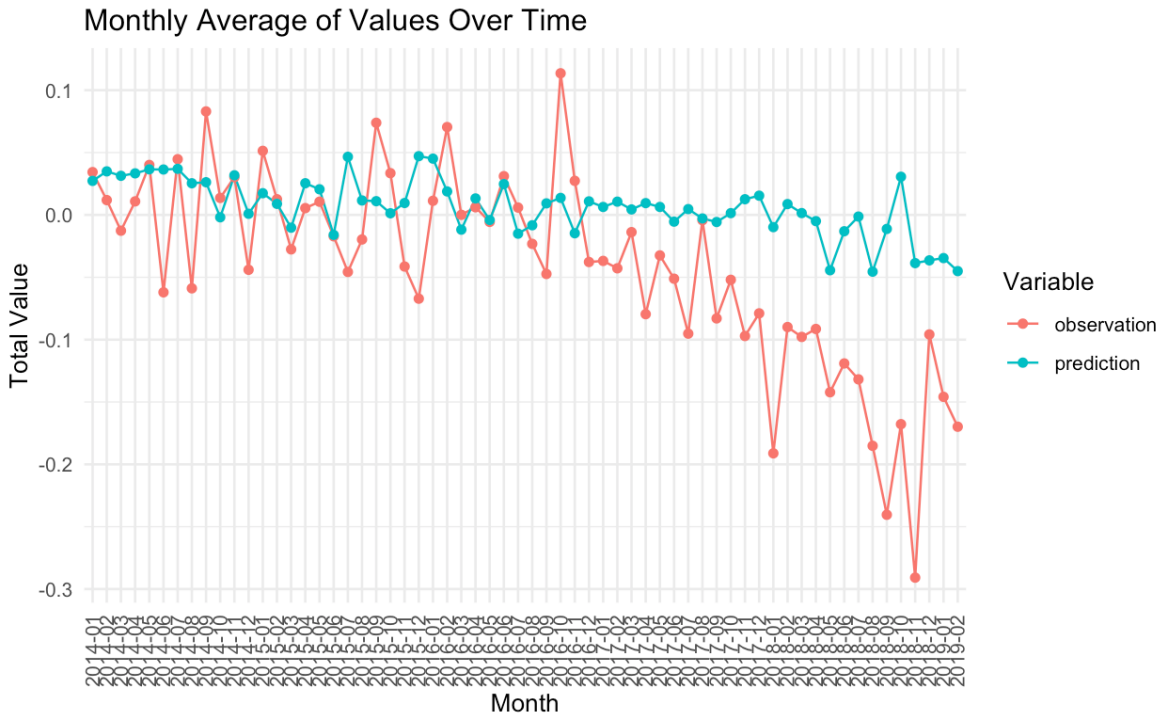
Table 1: Summary statistics for model performance

Model	MSE	MAE	Hit Ratio
Linear Regression	0.0268	0.0924	0.6443
Random Forest	0.0262	0.0907	0.6417
Support Vector Regression	0.0255	0.0881	0.6474

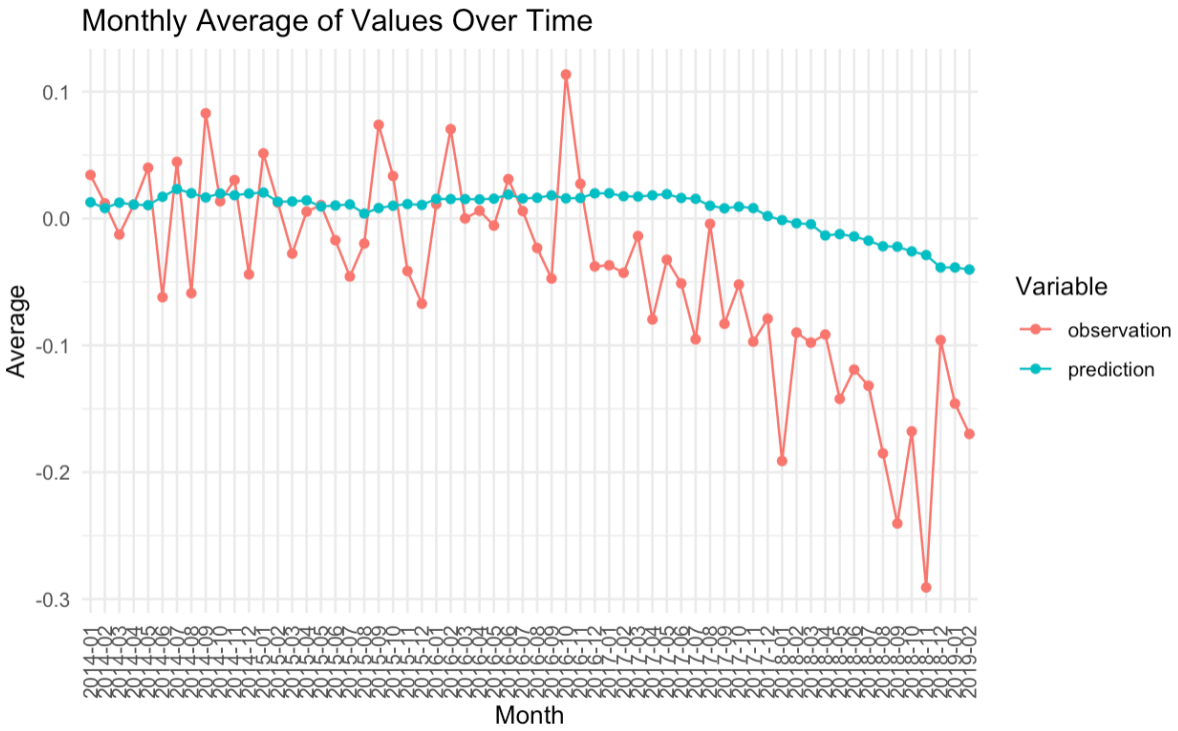
- *Linear Regression*: shows a moderate predictive capability as MSE of 0.0268, MAE of 0.0924 and Hit Ratio of 0.6443.
- *Random Forest*: displays an improvement over linear regression with a lower MSE and MAE. Its Hit Ratio is slightly lower than that of linear regression but remains close.
- *Support Vector Regression*: demonstrates the best performance among these three models with the lowest MSE of 0.0255, the lowest MAE of 0.0881, and the highest Hit Ratio of 0.6474. These performance statistics suggest that SVR is the most accurate in predicting both the magnitude and direction of stock returns.

Compared to these three models, the lower MSE and MAE of machine learning models imply an advantage in predicting magnitude of stock returns. This finding supports the hypothesis that machine learning models outperform linear regression due to their ability to capture non-linear relationships. However, the similar Hit Ratios indicate the predication of direction of stock returns is relatively consistent across models. Future studies will focus on improving the Hit Ratio brought by machine learning models.

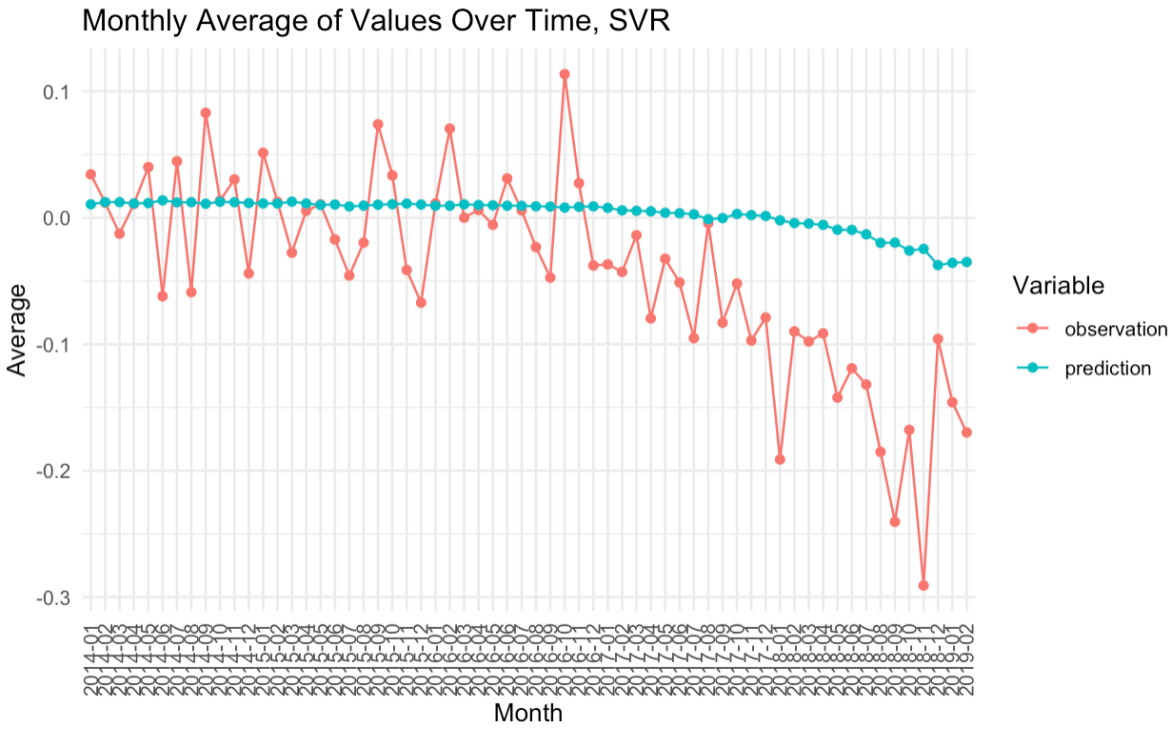
Graph 1: Average returns, predicted vs actual, from Feb 2014 to Mar 2019, linear regression



Graph 2: Average returns, predicted vs actual, from Feb 2014 to Mar 2019, RF



Graph 3: Average returns, predicted vs actual, from Feb 2014 to Mar 2019, SVR



In *Graph 1*, there exists significant extreme values, particularly when the observed values reach a relative peak. The predictions often show an opposite peak, as seen in July 2015 and December 2015. This indicates that the linear regression model may have trouble capturing the peaks and troughs in the data, possibly due to its inability to model the complex and non-linear relationships present in current financial markets.

Conversely, in *Graph 2* and *Graph 3*, the predictions from Random Forest (RF) and Support Vector Regression (SVR) generally follow the trend of the observations, such as showing a gradual decline since 2017, and their predictions show lower volatility than the observations, especially in the case of SVR. This is consistent with the characteristics of machine learning models that avoid overfitting to noise. Compared to the three graphs, it is evident that the

machine learning models outperform linear regression, in terms of closeness to actual values and smoothness of predictions.

However, while the predictions from RF reflect the observed volatility to some degree, they still exhibit relatively low volatility. All three models demonstrate challenges in accurately representing volatility. The linear regression model tends to overestimate volatility, whereas the machine learning algorithms tend to underestimate it, indicating a need for a balance between overfitting and underfitting. This will be investigated in future studies.

6. Related work

Projects 1 and 2 are highly related, therefore the reader is referred to our GPR#1 report for related work.

7. Conclusion

The comparative analysis in this project has demonstrated the outstanding performance of machine learning models within the Fama-French Five-Factor Model. This outperformance by machine learning algorithms not only indicates the promise of machine learning application in asset pricing models but also suggests the presence of non-linear relationships between the Fama-French Five Factors and stock returns.

These findings support innovation in the financial market, particularly in quantitative investing. Investment analysts can leverage the advantages of machine learning algorithms in asset pricing models for alpha seeking, portfolio optimization, risk management, and asset allocation.

Future work should continue to extend machine learning models into the Fama-French models and other asset pricing models. It aims to achieve more accurate predictions and a deeper understanding of financial market dynamics through the integration of machine learning algorithms and financial data analysis.

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