Financial Data Professional Institute
A Conversation with Joe Simonian

A Machine Learning Approach to Risk Factors: A Case Study Using the Fama-French-Carhart Model

Mehrzad Mahdavi, Executive Director, FDP Institute
Kathy Wilkens, Senior Advisor, FDPI Curriculum
Mirjam Dekker, Project Manager, FDP Institute

www.fdpinstitute.org
February 27, 2020
Agenda

• Welcome
• Introductions

Joe Simonian, Founder and CIO, Autonomous Investment Technologies, LLC; Co-editor of The Journal of Financial Data Science
Katherine Wilkens, Sr. Curriculum Advisor FDPI
Mehrzad Mahdavi, Executive Director FDPI
Mirjam Dekker, Project Manager FDPI

• Fama-French-Carhart Model
• Drawbacks of Traditional Quant Toolkit
• Pros of Applying ML in Finance
• Case Study
• FDP Curriculum
Drawbacks of the Traditional Quant Toolkit

The basis of much of contemporary risk measurement and investment management is based on Ordinary Least Squares (OLS) regression, developed by Carl Friedrich Gauss more than two centuries ago. Nevertheless, it possesses a number of shortcomings that make it a less than ideal framework for analyzing financial data characterized by:

- Non-linear threshold relationships
- Variables which exhibit important dependence relationships
- The interaction between numerical and categorical variables
- Other statistical and econometric models possess many of these shortcomings (and others), e.g. PCA.

Key takeaway: Investment practitioners current econometric toolkit is inadequate for dealing with the nuances of financial data sets

Joseph Simonian, Ph.D.
Founder and CIO, AUTONOMOUS INVESTMENT TECHNOLOGIES LLC
Co-editor of The Journal of Financial Data Science
Pros of Applying ML to Investing

Machine Learning frameworks gain much of their problem-solving power from the unique ways in which they process information. For example:

- Hierarchical approaches such as Random Forests
- Inherent parallel processing such as that found in Neural Networks
- Visual/graphical modes of analysis such as that utilized in Spectral Clustering

The variety of ML methods for processing information is invaluable in finance. For example, the risk management of multi-asset portfolios often involves uncovering and monitoring indirect relationships (e.g. between credit and equity) and detecting the likelihood of tail events that could have cross-asset impact. It often takes more than one approach to discover these relationships.

Key takeaway: Machine learning methods provide the technical means to conduct more nuanced and complete investment analysis

Joseph Simonian, Ph.D.
Founder and CIO, AUTONOMOUS INVESTMENT TECHNOLOGIES LLC
Co-editor of The Journal of Financial Data Science
Case Study: ML Meets the Fama-French-Carhart Model

Fama-French-Carhart Factor Specification

\[
E[R_s] = r_f + \beta_s^{Mkt} (E[R_{Mkt}] - r_f) + \beta_s^{SMB} E[R_{SMB}]
+ \beta_s^{HML} E[R_{HML}] + \beta_s^{PR1YR} E[R_{PR1YR}]
\]

Mkt: Market Beta

SMB captures Size Effect: Stocks with lower market capitalizations have been found to have higher average returns.

HML captures Value: Stocks with low book-to-market ratio have historically produced positive risk-adjusted returns.

PR1YR captures Momentum: Buying stocks that have had past high returns and shorting stocks that have had past low returns has historically produced higher average returns.

Joseph Simonian, Ph.D.
Founder and CIO, AUTONOMOUS INVESTMENT TECHNOLOGIES LLC
Co-editor of The Journal of Financial Data Science
Case Study: Random Forests

Decision Nodes

Root Node
GDP > 3%

Fact Nodes

Inflation > 4%

Raise Fed Funds Rate

Current Fed Funds < 1%

Raise Fed Funds Rate

Do Not Raise Fed Funds Rate

Do Not Raise Fed Funds Rate

Joseph Simonian, Ph.D.
Founder and CIO, AUTONOMOUS INVESTMENT TECHNOLOGIES LLC
Co-editor of The Journal of Financial Data Science
### Factor Percentile Returns and Equity Sector Predicted Values (monthly returns, Jan 1991 to Aug 2018)

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Consumer Discretionary</th>
<th>Rm-Rf</th>
<th>SMB</th>
<th>HML</th>
<th>MoM</th>
<th>Consumer Staples</th>
<th>Rm-Rf</th>
<th>SMB</th>
<th>HML</th>
<th>MoM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10th</td>
<td>-5.2%</td>
<td>-4.7%</td>
<td>-3.3%</td>
<td>-3.0%</td>
<td>-4.7%</td>
<td>-3.3%</td>
<td>-12.4%</td>
<td>-10.8%</td>
<td>-7.6%</td>
<td>-21.2%</td>
</tr>
<tr>
<td>25th</td>
<td>-1.7%</td>
<td>-1.9%</td>
<td>-1.8%</td>
<td>-1.5%</td>
<td>-1.4%</td>
<td>-0.9%</td>
<td>-5.3%</td>
<td>-3.6%</td>
<td>-3.6%</td>
<td>-5.4%</td>
</tr>
<tr>
<td>50th</td>
<td>1.3%</td>
<td>1.4%</td>
<td>0.3%</td>
<td>0.2%</td>
<td>1.0%</td>
<td>1.0%</td>
<td>2.2%</td>
<td>1.0%</td>
<td>1.0%</td>
<td>1.9%</td>
</tr>
<tr>
<td>75th</td>
<td>4.4%</td>
<td>3.7%</td>
<td>2.4%</td>
<td>2.1%</td>
<td>3.3%</td>
<td>2.9%</td>
<td>5.1%</td>
<td>3.4%</td>
<td>3.4%</td>
<td>4.6%</td>
</tr>
<tr>
<td>90th</td>
<td>6.7%</td>
<td>5.9%</td>
<td>3.9%</td>
<td>3.9%</td>
<td>5.2%</td>
<td>4.1%</td>
<td>11.4%</td>
<td>22.1%</td>
<td>12.9%</td>
<td>18.3%</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS $R^2$</td>
<td>0.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Energy</th>
<th>Rm-Rf</th>
<th>SMB</th>
<th>HML</th>
<th>MoM</th>
<th>Financials</th>
<th>Rm-Rf</th>
<th>SMB</th>
<th>HML</th>
<th>MoM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10th</td>
<td>-4.6%</td>
<td>-4.5%</td>
<td>-3.3%</td>
<td>-2.9%</td>
<td>-4.2%</td>
<td>-5.9%</td>
<td>-3.2%</td>
<td>-2.8%</td>
<td>-2.4%</td>
<td>-2.9%</td>
</tr>
<tr>
<td>25th</td>
<td>-2.1%</td>
<td>-1.2%</td>
<td>-1.3%</td>
<td>-1.1%</td>
<td>-0.8%</td>
<td>-2.4%</td>
<td>-1.2%</td>
<td>-1.2%</td>
<td>-1.1%</td>
<td>-0.6%</td>
</tr>
<tr>
<td>50th</td>
<td>0.8%</td>
<td>0.8%</td>
<td>-0.2%</td>
<td>-0.3%</td>
<td>0.2%</td>
<td>1.6%</td>
<td>1.4%</td>
<td>0.3%</td>
<td>0.1%</td>
<td>0.9%</td>
</tr>
<tr>
<td>75th</td>
<td>4.4%</td>
<td>3.8%</td>
<td>2.5%</td>
<td>2.2%</td>
<td>3.4%</td>
<td>4.1%</td>
<td>3.9%</td>
<td>2.5%</td>
<td>2.3%</td>
<td>3.5%</td>
</tr>
<tr>
<td>90th</td>
<td>6.5%</td>
<td>7.0%</td>
<td>5.0%</td>
<td>5.5%</td>
<td>7.0%</td>
<td>7.9%</td>
<td>4.8%</td>
<td>3.2%</td>
<td>3.2%</td>
<td>4.1%</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS $R^2$</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


---

**Joseph Simonian, Ph.D.**

Founder and CIO, AUTONOMOUS INVESTMENT TECHNOLOGIES LLC

Co-editor of *The Journal of Financial Data Science*
## Case Study: Sample Output

### Sector Rotation Strategy Backtest and Bootstrap Results (Jan 1997 to Aug 2018)

<table>
<thead>
<tr>
<th></th>
<th>Annualized Return</th>
<th>Annualized Volatility</th>
<th>PSharpe 0.0</th>
<th>PSharpe 0.1</th>
<th>PSharpe 0.2</th>
<th>PIR 0.0</th>
<th>PIR 0.1</th>
<th>PIR 0.2</th>
<th>Turnover</th>
<th>Total Cumulative Return</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Out-of-Sample Backtest Results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal-Weight</td>
<td>10.0%</td>
<td>13.8%</td>
<td>99.2%</td>
<td>78.8%</td>
<td>23.4%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0%</td>
<td>632%</td>
</tr>
<tr>
<td>Unconstrained Strategy</td>
<td>12.9%</td>
<td>14.2%</td>
<td>99.9%</td>
<td>94.9%</td>
<td>53.8%</td>
<td>96.9%</td>
<td>60.3%</td>
<td>10.3%</td>
<td>74.6%</td>
<td>1106%</td>
</tr>
<tr>
<td>Constrained Strategy</td>
<td>11.0%</td>
<td>13.6%</td>
<td>99.7%</td>
<td>87.3%</td>
<td>35.0%</td>
<td>97.8%</td>
<td>66.1%</td>
<td>13.6%</td>
<td>22.4%</td>
<td>773%</td>
</tr>
<tr>
<td><strong>Panel B: Bootstrap Performance Results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equal-Weight</td>
<td>11.0%</td>
<td>13.4%</td>
<td>96.7%</td>
<td>78.9%</td>
<td>43.0%</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0%</td>
<td>982%</td>
</tr>
<tr>
<td>Unconstrained Strategy</td>
<td>12.2%</td>
<td>14.4%</td>
<td>97.6%</td>
<td>82.2%</td>
<td>45.8%</td>
<td>66.7%</td>
<td>25.7%</td>
<td>4.2%</td>
<td>85.4%</td>
<td>1202%</td>
</tr>
<tr>
<td>Constrained Strategy</td>
<td>11.4%</td>
<td>13.4%</td>
<td>97.3%</td>
<td>81.3%</td>
<td>45.9%</td>
<td>66.7%</td>
<td>25.8%</td>
<td>4.3%</td>
<td>25.6%</td>
<td>1041%</td>
</tr>
</tbody>
</table>

FDP Curriculum & Ethics

1. Introduction to Data Science & Big Data
2. DM & ML: Introduction
3. DM & ML: Regression, LASSO, Predictive Models, Time Series & Tree Models
4. DM & ML: Classification & Clustering
5. DM & ML: Performance Evaluation, Backtesting & False Discoveries
6. DM & ML: Representing & Mining Text
7. Big Data, DM & ML: Ethical & Privacy Issues
8. Big Data and Machine Learning in the Financial Industry

Sample of the Reading(s):


Sample Keywords (of the Simonian reading):

<table>
<thead>
<tr>
<th>Factors (p. 32)</th>
<th>CART (p. 34)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear (p. 32)</td>
<td>Binary recursive partitioning (p. 34)</td>
</tr>
<tr>
<td>Nonlinear (p. 32)</td>
<td>Bagging (p. 34)</td>
</tr>
<tr>
<td>Random forest (p. 33)</td>
<td>Out-of-bag data (p. 34)</td>
</tr>
<tr>
<td>Supervised (p. 34)</td>
<td>Feature importance (p. 35)</td>
</tr>
<tr>
<td>Unsupervised (p. 34)</td>
<td>Mean decrease accuracy (p. 34)</td>
</tr>
<tr>
<td>Root node (p. 34)</td>
<td>Fama–French–Carhart (p. 37)</td>
</tr>
<tr>
<td>Decision node (p. 34)</td>
<td>Probabilistic Sharpe ratio (p. 42)</td>
</tr>
<tr>
<td>Terminal node (p. 34)</td>
<td></td>
</tr>
</tbody>
</table>
Sample Learning Objectives

8.4.1 Applications of random forest regression algorithm to factor models
For example:
A. Discuss two shortcomings of parametric nonlinear factor models that are developed to address shortcomings of linear models.
B. Discuss the ability of random forest algorithm to overcome one shortcoming of linear models.
C. Discuss the ability of random forest algorithm to overcome one shortcoming of parametric nonlinear models.
D. List 4 components of the decision tree when applied to the regression problem of factor models.
E. Describe how bagging is used in an ensemble of decision trees (random forest).
F. Calculate the predicted value of an dependent (response) variable given a set of predictor values and the outputs of a binary regression decision tree algorithm.
G. Describe the role of out-of-bag observations in a random forest algorithm.
H. Discuss mean decrease accuracy approach to estimating feature importance in a random forest algorithm.
I. Recognize and apply the probabilistic Sharpe ratio.

Sample Questions:
According to “A Machine Learning Approach to Risk Factors,” by Simonian et al., what is one of the shortcomings of non-linear factor models in comparison to linear factor models?
   a) The impacts of correlated factors are difficult to measure
   b) The parameter estimates cannot always be derived analytically
   c) The number of factors that must be used is relatively large

Source: FDP Institute Study guide March 2020 Exam
Source: LO 8.4.1, Reading 8.4, pp 32-33
Q & A

Kind reminders of upcoming webinars as we go through the Q & A. Add your questions in the chat room please.
In Closing

- The Next FDP Exam: March 16 –April 4, 2020
- Registration is open
- For a recent candidate webinar go to [www.fdpinstitute.org/webinars](http://www.fdpinstitute.org/webinars)

Learn more about the FDP Institute at [www.fdpinstitute.org](http://www.fdpinstitute.org)