Financial Data Analytics Report

CFM301 Portfolio Analysis

Sean Lee

Germain Zhang-Houle

Rain Luo

Jack Zhao

Alan Zheng

Ballerina Liang

**April 23, 2025**

**University of Waterloo**

**CFM 301**

**Introduction & Motivation**

Factor-based investing is one of the most influential and popular approaches in asset management. By finding key return drivers, investors can create portfolios that outperform the risk standard. The objective of this project is to design and evaluate a multi-factor equity strategy. For this project, the S&P 1500 stocks make up our investment universe, consisting of monthly data between 1980 and 2019. By running the Fama-MacBeth cross-sectional regression, we identify the factors that hint at a stronger predictive power through estimating risk premia. We then test these factors’ performance in a long-short portfolio. The training data spans the first 75% of the dataset (1980 - 2009), while the model is tested on the remaining 25% (2010 - 2019) to evaluate its efficacy in the real world. Through this project, we aim to test the hypothesis that a well-constructed, statistically validated set of factors can provide persistent return signals that outperform traditional benchmarks on a risk-adjusted basis.

**Data Processing & Cleanup**

To ensure the quality of our model, we conducted a series of data preprocessing steps. To start, we applied basic screening to remove unrepresentative stocks. More specifically, we filtered out:

* Stocks with a January share price below $5, and
* Stocks with a market capitalization of less than $100 million.

This step makes sure that our strategy targets only liquid, tradable stocks, which avoids biases from penny stocks or micro-caps. To handle missing data, we utilized cross-sectional median imputation. For each factor in each month, the missing values were replaced with the median value across all stocks in that month. Median imputation was selected because it is more robust to outliers than mean imputation, preserving the integrity of the cross-sectional distribution.

Next, all factor variables were winsorized to reduce the impact of outliers on our results. Extreme values can skew both regression results and portfolio weights, potentially distorting factor loadings and creating misleading signals. After winsorization, we standardized all factors within each month using z-scores. This allows for better comparability between factors and stocks since each value is expressed relative to the monthly cross-sectional mean and standard deviation. Overall, preprocessing the data confirmed that the data was clean, consistent, and ready to be regressed.

**Factor Selection**

To assess the predictive strength of each of the 50 provided factors, we first applied the Fama-MacBeth two-pass regression on the training data. The initial Fama-Macbeth regression yielded t-statistics in appendix **Exhibit A**. The first pass computed the betas while the second pass estimated the average risk premium and the associated t-statistics for each factor. We then ranked these 50 factors based on the absolute value of their t-statistics, with factors like “MoneyFlowIndex\_120\_z“ and “xret\_indsize\_std20\_z” showing stronger signals. We selected a subset of 9 factors to proceed with.

Using this selection, we screened for excessive collinearity between factors. We ultimately decided on the data science standard of a correlation of 0.7 being the limit. Pair correlations between factors were examined by taking the mean of correlation coefficients between two factors across all stocks; i.e. we calculated the correlation between the factors while aggregating by stock, then averaged the matrices (**Exhibit B**). This method accounts for the hierarchical nature of the data, and ensures that excessive collinearity is not obscured. We noticed the following:

* “IV\_capm\_z”, or **Idiosyncratic Volatility per the CAPM** Model, was highly correlated with both “xret\_indsize\_std20\_z” (**Standard Deviation of Industry and Size adjusted returns**) and “mdr\_z” (**Maximum Daily Return for 1-month holding period**) (ρ > 0.7).
* xret\_20\_z and xret\_indsize\_20\_z were strongly correlated (ρ > 0.7)
* range\_20\_z had a high correlation with volatility-related factors already included

To reduce redundancy and also improve model transparency, we removed factors that were less statistically significant in each collinear group. More specifically, we dropped IV\_capm\_z, xret\_20\_z, and range\_20\_z. After filtering for both t-statistic significance and collinearity, we arrived at the following six factors for our final model:

| **Factor Name** | **Description** |
| --- | --- |
| log\_vol\_dollar\_120\_z | 120-day log of dollar trading volume |
| BM\_z | Book-to\_market ratio |
| mdr\_z | Maximum daily return over the past month |
| xret\_indsize\_std20\_z | Std. dev. of 20-day excess returns |
| MoneyFlowIndex\_120\_z | 120-day money flow index |
| xret\_indsize\_20\_z | 20-day excess return adjusted for industry |

**Specific Factor Explanations**  
1. 120-Day Log of Dollar Trading Volume (log\_vol\_dollar\_120\_z)

This factor serves as a potential explanatory factor because of its association with market liquidity and participation levels. High-volume stocks undergo more transactions which shows strong investor attention and quicker integration of information into stock prices. This dynamic decreases mispricing risks while boosting how prices accurately reflect information. Stocks that experience low trading activity often go unnoticed by investors which results in delayed information dissemination and unusual return patterns. The logarithm of dollar volume both balances extreme figures and normalizes size differences between firms which results in a more accurate measure for diverse samples. Empirical research reinforces this rationale. The study by Brennan, Chordia and Subrahmanyam (1998) reveals that trading volume provides valuable insights into anticipated stock returns because it represents both market liquidity and levels of investor attention which are fundamental factors in asset pricing.

2. Book-to-Market ratio (BM)

This factor stands as an established metric for determining a firm's value and successfully predicts differences in stock returns across various sectors. Value stocks which demonstrate high B/M ratios usually show market prices that fall below their book values suggesting potential undervaluation or lack of investor favor. The potential for higher risk and behavioral biases among investors can make these firms produce greater expected returns. While low B/M stocks known as “growth” stocks are typically priced based on optimistic future expectations they often produce lower realized returns. The influence of this factor is well-established in empirical finance research with Fama and French (1992) identifying the B/M ratio as an essential variable in their asset pricing three-factor model. The research demonstrates that the B/M effect maintains its predictive power across different time periods and markets while serving as a fundamental indicator of expected equity returns.

3. Maximum Daily Return over 1-month period (mdr)

This factor serves as an indicator for both investor sentiment and speculative behavior. Retail traders and momentum-focused funds often direct their attention toward stocks that record substantial single-day gains which subsequently boosts demand and places upward pressure on prices in the following timeframes. Investor overreaction to significant signals generates this attention-driven behavior which results in stocks with lottery-like features maintaining high short-term returns through delayed price corrections and herd behavior. Investor preference for positively skewed returns as supported by Kumar (2009) leads to sustained demand for stocks generating extreme positive daily returns and positive return continuation. Market sentiment has a significant influence on pricing which creates a positive association between maximum daily returns and future returns.

4. Standard Deviation of 20-day Excess Returns (xret\_indsize\_std20)

This factor functions as a measure of short-term market volatility while also conveying information about company-specific risk perceptions. When companies exhibit high short-term volatility this signals greater uncertainty about future cash flows and operational stability which leads risk-averse investors to seek higher expected returns as compensation for increased risk. Investors perceive higher future returns possibilities when they link greater volatility with temporary fundamental risks instead of structural problems. Ang et al. Ang et al. (2006) found higher idiosyncratic volatility leads to lower future stock returns but later research by Fu (2009) showed that using conditional volatility models volatility can predict positive returns. The standard deviation of short-term returns may forecast future returns positively when market participants first overrate short-term volatility risks before market prices adjust upwards as uncertainty dissipates.

5. 120 Day Money Flow Index (MoneyFlowIndex\_120)

This factor integrates price and volume information to measure both buying and selling forces across a medium-term period. A continuously elevated MFI level indicates robust institutional buying activity or broad investor confidence which may signal future positive returns. This interpretation suggests that continuous inflows usually indicate upcoming positive corporate outcomes or unexpected earnings results. The use of volume as part of its calculation allows the Money Flow Index (MFI) to function differently from typical indicators based only on price by increasing its responsiveness to market conviction behind price movements. Studies have shown that both fund managers and institutional traders use technical indicators such as MFI when making investment decisions in markets where fundamental information is scarce (Menkhoff, 2010). The MFI's ability to detect hidden information and trading patterns enhances its importance for predicting stock returns because these patterns haven't yet influenced price movements.

6. 20-day Industry-Adjusted Excess Return (xret\_indsize\_20\_z)

This factor aims to identify company-specific momentum by accounting for general sector movements. When a company's stock surpasses its industry competitors for several consecutive periods it indicates potential unique positive developments the firm has experienced or reveals enhanced operational effectiveness along with strong investor confidence. Financial literature extensively documents momentum effects and Jegadeesh and Titman (1993) demonstrated that stocks which performed well in the past continue their strong performance in the short term. The factor refines stock-level alpha measurement by removing industry-wide shocks and macroeconomic effects through industry performance adjustment. The adjusted measure proves especially valuable for capturing cross-sectional return variation because companies displaying robust relative momentum typically stand out as leaders in innovation or operational execution.

**Methodologies Used & Results**

We used the Fama-MacBeth (1973) regression method to analyze the relationship both to choose a final factor set, as well as measure the risk premia associated with each final factor. After selecting and initially evaluating each factor through the aforementioned vetting process, we simulated predicted returns for the chosen firms during the out-of-sample period from 2010 to 2019. This was done using a custom model for each stock for each month, with a 6-year lagging window consisting of at most 72 data points.

In order to test the out-of-sample predictive power of the model, we backtested a long-short equity portfolio over the course of the out-of-sample period, rebalanced every month, with performance based on the predicted returns. Each month, we rebalance the portfolio to hold equal-weight long positions on the top ten stocks, alongside an equal-weight short position on the bottom ten stocks. This portfolio was observed for one month, after which it would be rebalanced every month. Unfortunately, the portfolio results were underwhelming, with an annualized raw return of 5.43% and Sharpe ratio of -0.89, implying reward relative to volatility was negative.

Our hedge portfolio did underperform. In CAPM, the portfolio suffered a negative monthly alpha -0.5822% (T-stat of -3.22). FN3 showed similar results with -0.5642% (T-stat of -3.06). R-squared values also remained quite low, 2.43%, CAPM with FF3 was 2.97%, meaning traditional risk factors only justify a bit of return variation. The annualized information ratio based on FF3 turns out to be -1.03, which means those portfolios tend to underperform.

**Conclusion**

The out-of-sample dataset used to validate our model covers over a decade's worth of market data from January 2010 and onwards; this is a relatively large timeframe that captures various market conditions that ultimately reveal the capacity of our predictive model. However, after computing the performance analytics on the hedge portfolio, we found that the model did not perform very well during the out-of-sample period, having produced negative annual returns (-5.42%) and statistically significant negative alpha (-0.5642%) as indicated by **Exhibit D**.

While occasional negative returns are to be expected and aren't particularly concerning in isolation, our model's consistent underperformance across the out-of-sample testing period seems to suggest that there are some issues with its predictive ability. To address these shortcomings we could try iteratively refining the model by revisiting the factor selection process, adjusting key parameters such as the rolling window size, or exploring alternative modeling approaches altogether.

As part of our factor selection process we strategically selected factors that demonstrated both strong statistical signals and low multicollinearity, however, rather than this heuristic approach, it is possible that a more systematic approach that involves regularizing to avoid overfitting such as using a Lasso regression would have produced a more robust set of factors.

It is also possible that the 6-year rolling window used in the current implementation is too long and overly diminishes recent information. Such a window was chosen to ensure more than 10 data points for each factor, to prevent overfitting. The low frequency continued to prove to be a challenge for us, as such a long lookback period risks acclimatizing our model to a past financial regime, mismatching it to the regime it is intended to predict. Furthermore, a factor model designed around factors based in technical analysis would theoretically better predict returns in a higher frequency model, reaping returns before market sentiment turns or price corrections occur. In the future, we believe there are a few measures we can take to mitigate such issues:

* Use higher-frequency data to decrease overfitting and increase the quality of our model.
* Use more complex functions to more accurately attribute factors to the movements of certain stocks.
* Use an exponentially weighted moving average (EWMA) scheme to weigh recent observations more heavily.

Other areas we may try differently in the future have to do with the model itself and using our own factors. Our current approach uses a fairly standard OLS regression with a rolling window to predict returns. However, it's possible that more advanced techniques that introduce a penalty element (such as Ridge and Lasso regressions) or machine learning models (such as elastic net and neural networks) could yield better results. This is because a theoretically profitable linear model would be easily replicable by other market participants, especially if it has persisted from the 1980-2009 era, making such a strategy unviable. From discussion amongst the group members we find that trying to come up with our own factors is very time consuming with the potential of not being statistically significant. We did agree that creating our own factors would not be a wise idea. However, given enough time, it would be interesting to create our own factors and see its correlation.

**Appendix**

**Exhibit A:** Initial Fama Macbeth Results

|  | **mean risk premia** | **Standard Error** | **T statistics** |
| --- | --- | --- | --- |
| **IM\_z** | 0.020039 | 0.022124 | 0.905764 |
| **range\_20\_z** | -0.01483 | 0.010598 | -1.399315 |
| **log\_vol\_dollar\_20\_z** | 0.000251 | 0.004533 | 0.055287 |
| **range\_120\_z** | 0.002939 | 0.004945 | 0.59438 |
| **log\_vol\_dollar\_120\_z** | 0.002822 | 0.002657 | 1.06232 |
| **xret\_5\_z** | -0.01764 | 0.031591 | -0.558397 |
| **xret\_10\_z** | -0.031334 | 0.029885 | -1.048456 |
| **xret\_20\_z** | -0.036119 | 0.025483 | -1.417355 |
| **xret\_indsize\_20\_z** | -0.027634 | 0.026518 | -1.042086 |
| **xret\_indsize\_std20\_z** | -0.02371 | 0.014865 | -1.595028 |
| **xret\_40\_z** | -0.01729 | 0.028016 | -0.61717 |
| **xret\_120\_z** | -0.003333 | 0.016948 | -0.196649 |
| **xret\_indsize\_120\_z** | -0.004976 | 0.016668 | -0.298556 |
| **xret\_indsize\_std120\_z** | -0.005344 | 0.006258 | -0.853913 |
| **KDJ\_20\_z** | -0.033601 | 0.040575 | -0.828125 |
| **deviation\_pct20\_z** | -0.019277 | 0.028469 | -0.677121 |
| **MoneyFlowIndex\_20\_z** | 0.011897 | 0.034693 | 0.342926 |
| **RSI\_20\_z** | -0.030869 | 0.032609 | -0.946653 |
| **KDJ\_120\_z** | 0.006708 | 0.017275 | 0.388328 |
| **deviation\_pct120\_z** | -0.011338 | 0.018253 | -0.621123 |
| **MoneyFlowIndex\_120\_z** | 0.033797 | 0.020561 | 1.643756 |
| **RSI\_120\_z** | 0.019277 | 0.021492 | 0.896937 |
| **IV\_capm\_z** | -0.023295 | 0.01458 | -1.597701 |
| **mdr\_z** | -0.029483 | 0.020194 | -1.460001 |
| **ami\_3\_z** | -0.001167 | 0.002422 | -0.482053 |
| **beta\_3y\_z** | 0.001571 | 0.009304 | 0.168806 |
| **beta\_5y\_z** | -0.000563 | 0.006327 | -0.089054 |
| **tail\_2y\_z** | -0.017747 | 0.022018 | -0.806008 |
| **dp\_z** | 0.000806 | 0.003253 | 0.247601 |
| **leverage\_z** | -0.000035 | 0.002948 | -0.011902 |
| **BL\_z** | 0.00026 | 0.001034 | 0.251909 |
| **roe\_z** | -0.000157 | 0.002873 | -0.05472 |
| **roa\_z** | -0.003446 | 0.003349 | -1.028992 |
| **profitability\_z** | -0.00114 | 0.003306 | -0.344859 |
| **sales\_g\_q\_z** | 0.003416 | 0.007101 | 0.481003 |
| **sales\_g\_ttm\_z** | 0.003284 | 0.007945 | 0.413313 |
| **op\_income\_g\_q\_z** | 0.005901 | 0.007135 | 0.827085 |
| **ni\_g\_q\_z** | 0.00535 | 0.005838 | 0.916453 |
| **op\_income\_g\_ttm\_z** | 0.003522 | 0.004439 | 0.793527 |
| **ni\_g\_ttm\_z** | 0.00024 | 0.004248 | 0.056485 |
| **sue\_NI\_z** | 0.012622 | 0.013881 | 0.909238 |
| **BM\_z** | 0.004348 | 0.003369 | 1.290654 |
| **AM\_z** | 0.000263 | 0.002049 | 0.128166 |
| **EP\_z** | 0.001647 | 0.003058 | 0.538639 |
| **SP\_z** | -0.000744 | 0.002933 | -0.253619 |
| **roe\_q\_z** | 0.000882 | 0.004826 | 0.182773 |
| **roa\_q\_z** | 0.002565 | 0.006003 | 0.427271 |
| **Cto\_z** | 0.000882 | 0.002499 | 0.352904 |
| **pe\_ttm\_z** | -0.001768 | 0.002211 | -0.799637 |
| **lag\_log\_size\_z** | 0.000048 | 0.002666 | 0.01814 |

**Exhibit B:**

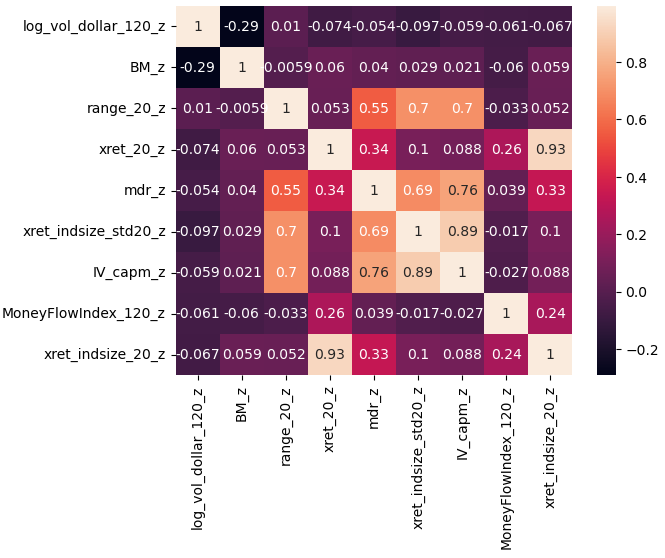
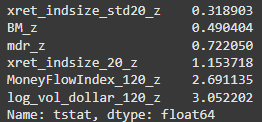
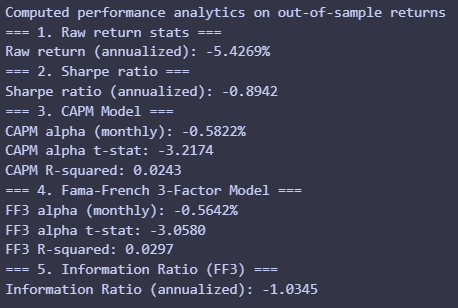


Exhibit B: Note the lighter cells, indicating high positive correlation. Negative correlation never reaches below -0.3, so there is no highly-negative correlation.

**Exhibit C:**



**Exhibit D:** Computed Performance Analytics on Resulting Out-of-sample Returns

****

**Project Requirements:**

The submitted repository contains a Jupyter Notebook file named “**CFM301\_Final\_Project.ipynb**” that includes the logic for our quantitative trading strategy.

The script requires the following Python libraries to run: **numpy, pandas, scipy, linearmodels, seaborn, statsmodel, and matplotlib.** “tqdm” is not necessary but highly recommended as a loading bar for our moving-window regression backtest.

The following data files to be located in the root directory:

* “**merged\_df.sas7dbat**” – the provided data set of 50 monthly factors for S&P 1500 stocks (S&P 500 + 400 + 600 stocks) from 1980 to 2019, alongside their monthly 1-month to 12-month-ahead returns.
* “**ff\_factors.xlsx**” – this data file includes the monthly Fama-French factor data used in the model evaluation section. It contains MKTRF (**Market Risk-Free Returns**), SMB (**“Small Minus Big” Factor**), HML (**“High Minus Low” Factor**), and RF (**“Risk Free Rate”**) spanning from 1980 to 2024.

**References**

Ang, Andrew, Robert J. Hodrick, Yuhang Xing, and Xiaoyan Zhang. The Cross‐Section of

Volatility and Expected Returns. *The Journal of Finance 61*, no. 1 (2006): 259–299. https://doi.org/10.1111/j.1540-6261.2006.00836.x

Brennan, M. J., Chordia, T., & Subrahmanyam, A. (1998). *Alternative factor specifications,*

*security characteristics, and the cross-section of expected stock returns*. Journal of Financial Economics, 49(3), 345–373. https://www.sciencedirect.com/science/article/abs/ pii/S0304405X98000282

Fama, Eugene F., and Kenneth R. French. The Cross‐Section of Expected Stock Returns.

*The Journal of Finance* 47, no. 2 (1992): 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Fu, Fangjian. Idiosyncratic Risk and the Cross-Section of Expected Stock Returns. J*ournal of*

*Financial Economics* 91, no. 1 (2009): 24–37. https://doi.org/10.1016/j.jfineco.2008.02.003

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications

for stock market efficiency. *The Journal of Finance*, 48(1), 65–91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Kumar, Alok. Who Gambles in the Stock Market? *The Journal of Finance* 64, no. 4 (2009):

1889–1933. https://doi.org/10.1111/j.1540-6261.2009.01483.x

Menkhoff, L. (2010). The use of technical analysis by fund managers: International evidence.

*Journal of Banking & Finance*, 34(11), 2573–2586. https://doi.org/10.1016/j.jbankfin.2010.04.014