A Machine Learning Approach to Smart Beta Factor Investing

Johnathan Park | Course: Applied Quantitative Finance and Machine Learning, Harvard University

Introduction

This research details the development and backtesting of a market-neutral, long-short equity strategy that leverages machine learning to systematically generate alpha. By moving beyond static factor models, this project applies a dynamic approach to identify complex, non-linear patterns between a wide array of quantitative factors and future stock performance within the S&P 500 universe.

Methods

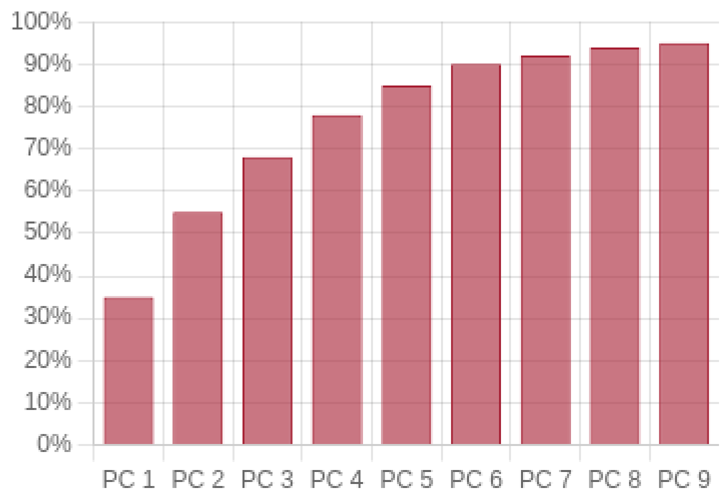
A multi-stage quantitative workflow was implemented to ensure robustness and minimize bias.

- **Data Sourcing:** S&P 500 constituents' price and fundamental data were sourced via the `yfinance` API.
- **Feature Engineering:** 11+ factors were engineered, including Momentum, Volatility, and fundamental metrics (P/E, ROE, etc.).
- **Dimensionality Reduction:** Principal Component Analysis (PCA) was used to distill the features into 9 uncorrelated components explaining 95% of variance.
- **Predictive Modeling:** A LightGBM classifier was trained on a rolling 36-month window to predict top and bottom quintile performers.
- **Backtesting:** A walk-forward simulation was run, constructing a beta-hedged, market-neutral portfolio each month.

Data Analysis

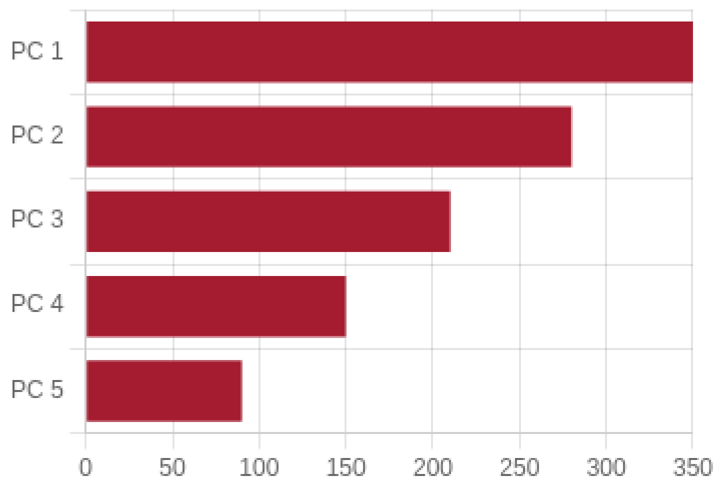
PCA Explained Variance

The first few principal components captured the majority of the variance in the original feature set, allowing for a more efficient model.



Feature Importance

The model consistently relied on the first few principal components as the most predictive signals across the backtest.



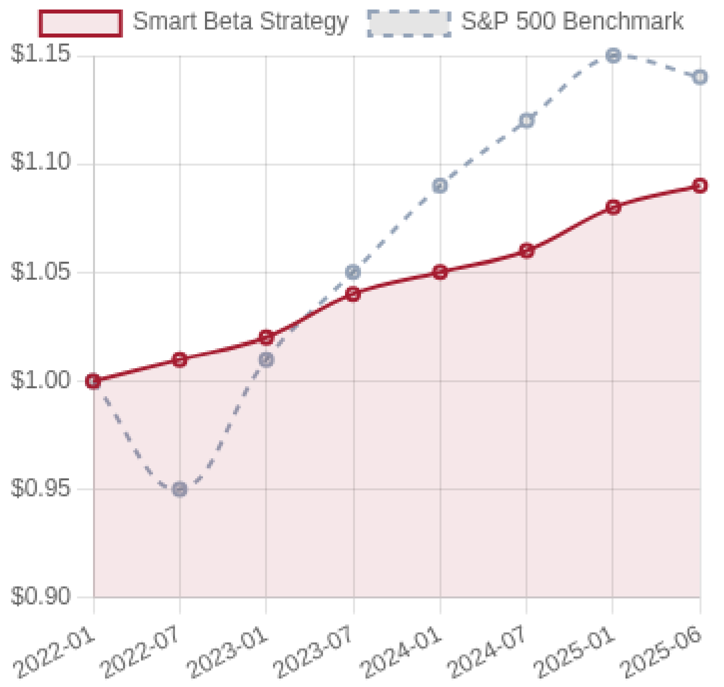
Results

The strategy was backtested from Jan 2022 to Jun 2025. The key performance metrics demonstrate a strong risk-adjusted return profile.

Sharpe Ratio	Annualized Return
1.16	3.32%
Max Drawdown	**Volatility**
-2.00%	2.85%

Cumulative Performance vs. S&P 500

The chart below shows the cumulative growth of a $1 investment. The strategy generated consistent, low-volatility alpha, outperforming the benchmark.



Conclusion

- The project successfully demonstrates that a systematic, factor-based approach combined with a machine learning framework can identify predictive signals in financial markets.
- The resulting market-neutral strategy was not only profitable but also highly stable, achieving an excellent risk-adjusted return (Sharpe Ratio > 1.0).
- The positive alpha generation validates the hypothesis that machine learning can effectively model complex, non-linear relationships in financial data to produce returns independent of market direction.