Stock Selection with Machine Learning: Featuring Analytical Techniques to beat the S&P 500 Benchmark

By Andrew Napurano
M.S. Data Science, Pace University
Advisor: Dr. Christelle Scharff

GitHub: anapurano211/Stock-Selection-with-Machine-Learning-CS668-Capstone-Project
Detailed Research : https://docs.google.com/presentation/d/1aSppbj_hVKVHL4m_4S7yzVH2PhDZKpjp/edit***Live***
Application: https://huggingface.co/spaces/andrewnap211/agentic-ai-financial-analyst

Introduction

Problem: 92% of active mutual funds underperform the benchmark (SPIVA).
Goal: Predict next-week company outperformance and build a machine learning based long-only portfolio.

- Empower investors with data-driven insights to outperform traditional benchmarks.
- Replace human bias and emotion with transparent, systematic decision-making.
- We build a weekly S&P 500 stock-selection pipeline using ~262k stock-week rows and 100+ features. We select 39 signals via Information Coefficient (IC) screening and evaluate models using rolling 3-year train / 1-year test windows.

Research Questions

Primary

1. Which fundamental + technical signals predict weekly outperformance?
2. Do drivers change by sector / regime (bull vs bear)?
3. Can a probability-ranked portfolio beat VOO on a risk-adjusted basis?

Secondary

1. Are the top signals stable year-to-year under rolling backtests?
2. Can the approach be deployed as an automated ETL + weekly rebalance pipeline?

Literature Review

Most equity stock-selection research relies on factor models and regressions to explain cross-sectional returns (Rasekhschaffe & Jones, 2019; Buczynski et al., 2021). Fewer studies apply ML classifiers directly to S&P 500 selection or test how factor importance shifts over time (Caparrini et al., 2024). Recent work benchmarks ML against regularized logistic regression (Wolff & Echterling, 2022) and uses value/quality features with ensembles like XGBoost to identify key predictors (Priel & Rokach, 2024). Overall, ML can capture nonlinear interactions among valuation, profitability, and momentum, but results often weaken under costs, robustness checks, and strict out-of-sample testing—making disciplined feature selection and realistic backtests essential (Buczynski et al., 2021).

Methodology

Feature Engineering & IC Screening

• Start from 100+ features; winsorize, standardize, and create composite factors.
• Compute historical Information Coefficient (Spearman correlation vs next-week returns) to rank signals.
• Freeze a 39-feature schema of IC-positive predictors (value, quality, momentum, regime flags).

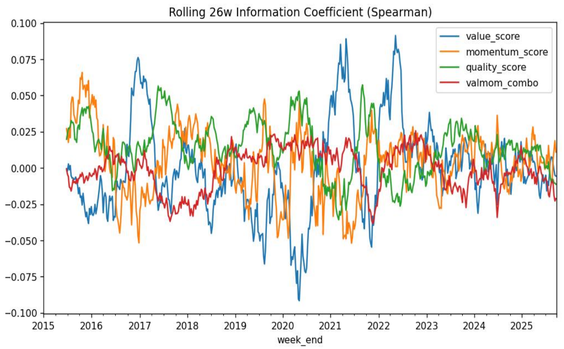
Rolling Train/Test Design

• Rolling 3-year train / 1-year test by calendar year (2015–2025).
• Train-only preprocessing to avoid look-ahead bias.
• Thresholds tuned per year to pick the top-probability bucket (e.g., top ~15–20%).

Model Zoo

• Logistic Regression (regularized)
• PCA + Logistic Regression
• Histogram Gradient Boosting
• Xtreme Gradient Boosting
• Decision Tree
• Stochastic Gradient Descent Linear Classifier
• Dense Deep Network

Information Coefficient Scores Overtime



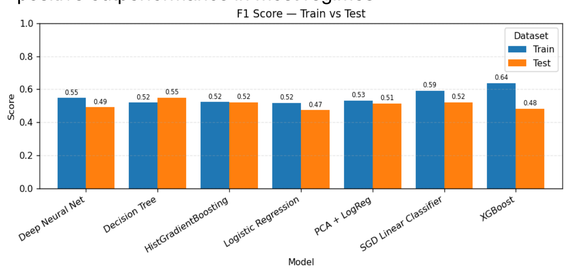
Results

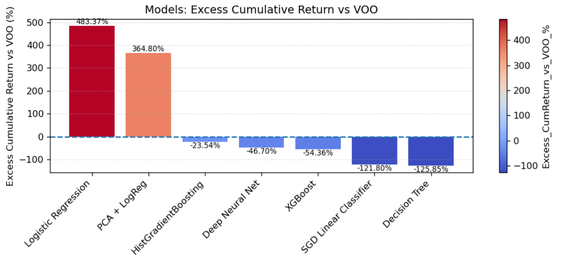
Model Performance vs VOO (All Test Years)

• Logistic Regression achieves 24.8% annualized return with 22.5% annualized volatility and a cumulative return of 7.68×, versus 14.8% annualized and 2.85× cumulative for VOO.
• PCA + Logistic Regression also outperforms with 22.9% annualized and 6.49× cumulative.
• Tree-based and deep models (HGB, XGBoost, deep net, SGD, Decision Tree) underperform VOO on both return and risk-adjusted basis.
•

Regime-Level Behavior

• The strategy strongly outperforms during the pre-COVID bull (+90.7% vs +40.3%) and post-COVID bull (+215.6% vs +89.2%).
• Performance is roughly in line during the COVID crash, and modestly worse in 2022 bear (−19.6% vs −15.4%).
• Post-2023, performance remains positive with smaller but still positive outperformance in most regimes





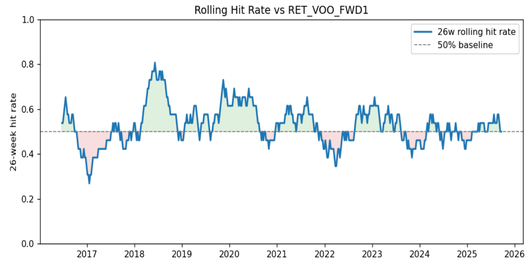
Conclusion

This work shows that a carefully engineered ML pipeline can turn noisy weekly S&P 500 data into a profitable stock-selection strategy. Regularized Logistic Regression using a compact set of IC-screened features delivers substantial outperformance versus VOO, both in total and within key market regimes,

Limitations
• Only one universe (S&P 500) and one weekly horizon.
• Transaction costs, slippage, and capacity constraints are not yet modeled.
• Hyperparameter tuning and architecture search for deep models are conservative.

Future Work
• Extend to multi-horizon (daily/monthly) labels and additional asset classes.
• Incorporate realistic trading frictions and turnover controls.
• Blend the ML alpha sleeve into a broader portfolio optimizer with regime-aware allocation.

Logistic Regression Model Hit Rate



References

1. Caparrini, A., Arroyo, J., & Escayola Mansilla, J. (2024). S&P 500 stock selection using ML classifiers. RIBAF, 70, 102336.

2. Wolff, D., & Echterling, F. (2022). Stock Picking with Machine Learning. SSRN Working Paper (No. 3607845).

3. Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine Learning for Stock Selection. Financial Analysts Journal, 75(3), 70–88.

4. Priel, R., & Rokach, L. (2024). ML-based stock picking with value + quality features. Neural Computing & Applications, 36(20), 11963–11986.

5. Buczynski, W., Cuzzolin, F., & Sahakian, B. (2021). Review of ML in equity investing (real-world gaps). IJDSA, 11(3), 221–242