Enhancing the Performance of Classic Investment Formulas with Machine Learning

Research Project for Finance

Abstract

This paper evaluates the application of Machine Learning methods with altered feature sets to four classic investment frameworks: the F-Score, Magic Formula, Acquirer's Multiple and Conservative Formula. All modified formula long only portfolios achieve higher raw, and risk adjusted returns, with lower drawdowns over the classic ones, and the market index. The F-Score and combined method achieve statistically significant FF5 + MOM alphas at the 1% and 5% level respectively. Each ML iteration achieves substantial cumulative returns over the original version for the testing window of 2010-2023, although only the F-Score, Magic Formula, and the total combination are significant at the 5% level. Post-hoc feature analysis reveals that machine learning is indeed useful for stock selection over a ranking system that accounts for metric scores equally. There are few predictors which drive most of the explanatory power within a set of formula features, and the relationships between them are not always linear. From each ML feature set, ROA, EY, Cash Flow-to-Price and NPY can be seen as most useful, leading to the highest probability of outperformance. While 3 out of 5 implementations are statistically robust, this study manages to address the main concerns of performance decay, as evidenced by the average improvement from classic combined to ML combined of 217.4% for the 14-year period.

Group 14 Team 1

Vasil Zhiliev (2815500)

Yunija Wan (2820650)

Bowen Sun (2835711)

Doortje van der Heide (2818319)

VU VRIJE UNIVERSITEIT AMSTERDAM

Table of Contents

1 Introduction

Systematic investment strategies, such as Piotroski's F-Score (Piotroski, 2000) Greenblatt's Magic Formula (Greenblatt, 2006), Carlisle's Acquirer's Multiple (Carlisle, 2017), and van Vliet and De Koning's Conservative Formula (van Vliet & Koning, 2017), have been widely adopted as frameworks for stock selection. These strategies are highly regarded in both academic and professional environments because due to their straightforward implementation, replicability, and ability to deliver outperformance. They have leveraged key factors such as value, momentum, and quality to outperform across time periods and in multiple geographies (Schwartz & Hanauer, 2024). This has established them as important and valuable tools for academics studying market efficiency and for professionals seeking superior risk-adjusted returns.

Despite their initial and historical success, these strategies can be seen to face challenges within the last ~10 years in the most recent literature. Empirical evidence shows that post publication, perhaps due to their widespread adoption, their effectiveness has been decreasing. While they were once able to generate a positive alpha, these strategies now seem to approach parity with the market index (Schwartz and Hanauer, 2024). Because these strategies are relatively static and rule-based in their nature, they are not suited to adapt to evolving and increasingly difficult market conditions. This decline in effectiveness has raised questions about the long-term viability of these strategies and simultaneously shows the importance of innovative solutions to sustain their relevance.

Machine learning (ML) presents a compelling opportunity to address the aforementioned limitations as they excel at combining multiple (weak) predictors into robust investing signals, which leads to outperformance in stock selection tasks, when compared to linear models (Gu et al., 2020). By overlaying the pre-existing framework of the classic formulas with ML models, a new set of formulas can be established. Through this application, each formula can dynamically adapt to market changes by uncovering non-linear relationships in returns and accounting data. Furthermore, the use of ML brings an added layer of sophistication and novelty over previous studies. It is highly relevant since additional (enhanced) accounting and economic variables can be added and grouped with the pre-existing features, in order to reflect their varying levels of utility within the entire model (Rasekhschaffe & Jones, 2019). These capabilities frame ML as a viable tool that modernizes traditional investment formulas by enhancing their adaptability and identifying both the individual and combined effects of features, all with the aim of maintaining their relevance in the rapidly evolving financial landscape.

This study aims to investigate the integration of ML techniques into the four prominent investment formulas – the Magic Formula (MF), Piotroski F-Score (F-Score), Acquirer's Multiple (AM), and the Conservative Formula (CF). The goal of the study is to:

1. Assess whether ML can mitigate performance decay observed in the traditional formulas.

2. Evaluate whether ML enhanced strategies achieve superior risk-adjusted returns.
3. Analyse the returns sources generated by ML models, to gain a more in depth understanding of the drivers of outperformance in modern markets.

By addressing the mentioned goals, this research aims to contribute to both academic literature and practical investing. It builds on existing studies on systematic investing and ML and aims to provide empirical evidence on how ML can enhance traditional investment frameworks. It offers insights for investors seeking new portfolio strategies capable of thriving in high-noise, complex, and dynamic markets. Furthermore, this study bridges the gap between traditional systematic investing and data-driven innovation. It provides a framework of sustained outperformance in contemporary financial markets.

In summary, this report explores the role of machine learning in the modernization of formula-based investment strategies, focusing on generating positive risk-adjusted returns. By doing so, it aims to provide a roadmap for enhancing systematic investing in an increasingly competitive and efficient financial landscape.

The main findings are a result of multiple evaluation criteria that can be summarised as follows:

Firstly, a holistic overview was graphed, showing the performance of each formula and its machine learning model, along with an equal weighted combination of all. The ML F-Score had the highest gross cumulative performance (714.7%), closely followed by the ML Magic Formula (592.8%). The combination of all formulas achieved a 217.4% increase over the original and 188.2% over the benchmark index.

Secondly, a detailed assessment of the period performance for all formulas along with significance tests for the cumulative outperformance for the ML versions over the original strategies is shown. All ML counterparts have notably better risk adjusted metrics with the biggest improvements again, stemming from the ML F-Score and Magic Formula, which offer an increase of 0.32 and 0.31 in terms of their Sharpe ratios, which is approximately 72% higher than the classic formulas' Sharpe ratios. All ML methods except for the ML CF have a decreased max drawdown which is purely resulting from the implementation of machine learning, as both classic and machine frameworks employed the same set of features and altered fundamental data, capitalized for R&D.

Thirdly, all formulas' returns are regressed onto a six-factor model (FF5+MOM). All formulas have a positive increase in alpha, however only that of the ML F-Score and equal weighted combination are significant at the 1%(!) and 5% levels respectively. Whatever level of returns increase is gained can therefore be attributed to a more adequate factor allocation arising from the machine learning models. None of the MF Formulas overweigh the market factor. Additionally, there is a trend across all

improvements that decreases the exposure to SMB while elevating HML and MOM scores. Besides the positive alpha scores, this is the focal point explaining the combined approach's outperformance.

Lastly, all features are analysed to gain a better understanding of the most influential ones, which implicitly carry the highest weight. We find that in the Acquirer's Multiple, the enhanced value measure-Cash Flow to Price- has more predictive power than the Earnings Multiple (EM) feature, although both equate to a higher probability of outperformance as they increase linearly. In contrast, the features of the Magic Formula appear to be non-monotonic. A stock with a high EY combined with low ROC has the same or a better probability of outperforming compared to a stock with a moderate to high ROC and a low EY.

2 Literature Review

2.1 Overview and Scope

Systematic investment strategies are the cornerstone of quantitative finance, offering formula-based frameworks that aim to outperform benchmarks such as the Capital Asset Pricing Model (CAPM). Notable ones include Greenblatt's Magic Formula (2006), Piotroski's F-Score (2000), Carlisle's Acquirer's Multiple (2017), and Van Vliet and Koning's Conservative Formula (2017). However, as mentioned, empirical evidence has shown that performance decay poses a significant challenge for these strategies, driven by their widespread adoption and evolving market conditions (Schwartz & Hanauer, 2024). ML has emerged as a promising tool to address these limitations by uncovering complex and non-linear relationships (Rasekhschaffe & Jones, 2019).

This literature review aims to provide an in-depth analysis behind the framework of the selected investment strategies. It first examines the theoretical performance of the selected strategies, emphasizing strengths and limitations. After this, it explores the capabilities of ML in systematic investing. Finally, it identifies critical gaps in the existing research and establishes the foundation for this study's investigation.

By contextualizing existing academic knowledge, this literature review sets the stage for evaluating whether ML can modernize traditional strategies and offer a significant edge in navigating the complexities of financial markets.

2.2 Investment Strategies

2.2.1 Piotroski F-Score

Piotroski (2000) developed the F-Score as a systematic investment strategy to evaluate the financial health of high book-to-market firms. It tries to identify healthy companies that have yet to fulfil their potential relative to their fundamentals, or that have been otherwise mispriced by the market. The scoring system utilizes nine binary financial signals to assess a company's profitability, leverage, liquidity, and operational efficiency. The total score is a sum of all binary indicators, and ranges from 0-9. Each criterion is assigned a 1 if its met and 0 otherwise. A higher total score indicates a stronger financial position and a higher likelihood of positive future returns. Companies forming a portfolio from the F-Score are the top decile with a score higher or equal to 7, sorted by ascending order in terms of their book to market ratio.

The nine components of the F-score can be divided into three categories:

1. Profitability
   - **Return on Assets (ROA):** Scores 1 if ROA in year *t* is positive.
   - **Operating Cash Flow (CROA):** Scores 1 if operating cash flow in year *t* is positive.
   - **Change in ROA (ΔROA):** Scores 1 if ROA in year *t* is higher than in year t-1
   - **Accruals (CROA - ROA):** Scores 1 if operating cash flow exceeds net income, indicating earnings quality.

2. Leverage, Liquidity, and Source of Funds:
   - **Change in Leverage (ΔLEV):** Scores 1 if the firm's leverage (long-term debt) has decreased or remained unchanged compared to the prior year.
   - **Change in Current Ratio (ΔLIQ):** Scores 1 if the current ratio (current assets divided by current liabilities) has improved year-over-year.
   - **Equity Issuance (EQIS):** Scores 1 if the firm has not issued new equity during the year, suggesting no dilution of existing shareholders.

3. Operational Efficiency:
   - **Change in Gross Margin (ΔGM):** Scores 1 if gross margin has improved from the previous year.
   - **Change in Asset Turnover (ΔTURNOVER):** Scores 1 if asset turnover (sales divided by total assets) has increased, indicating better utilization of assets.

Empirical evidence for the F-Score has shown robust support and effectiveness across markets. Walkshäusl (2020) found that high F-score firms significantly outperform low F-score firms by approximately 10% per year, after controlling for established return determinants. This result is supported by Deng (2016), who further explored this and found that this outperformance is particularly true for firms with low(er) book-to-market ratios. Furthermore, across European markets it was found that incorporation of the F-Score significantly improved the performance of existing value strategies, suggesting the ability to make a divide between 'winners' and 'losers' (Tikkanen & Äijö, 2018).

However, the dependence on binary signals might limit the effectiveness of the scoring system in capturing complex financial relationships. As markets evolve, the static nature of the scoring system may have a reduced predictive power, highlighting the need for machine learning enhancements.

2.2.2 Magic Formula

The Magic Formula is a systematic investment strategy developed by (Greenblatt, 2006). It is designed to identify undervalued yet high-performance stocks. The Magic Formula is based on two key financial metrics: Earnings Yield (EY) and the Return on Capital (ROC). EY calculates how much a company earns relative to its market value. ROC evaluates a company's efficiency in generating profits from operational investments. The Magic Formula aims for the stocks with the highest Magic Formula Rank (i.e. the highest combined EY and ROC rankings). These stocks are selected for inclusion in the portfolio as they display low earnings multiples while being operationally efficient.

While the main application was highlighted to be in US stocks, empirical studies have shown that the Magic Formula has the ability to outperform benchmarks over wide time periods and across different markets. Gunnar Juliao de Paula (2016) demonstrated that the Magic Formula has outperformed the Brazilian 'Ibovespa Index' while addressing a multitude of biases. This outperformance was later confirmed by Schwartz and Hanauer (2024) who showed significant risk-adjusted returns across markets. These findings confirm the Magic Formula's ability to effectively capture value. However, similarly to the other strategies, performance of the Magic Formula has shown signs of decline in the latter years of its cumulative performance (Schwartz & Hanauer, 2024).

2.2.3 Acquirer's Multiple

The Acquirer's Multiple, introduced by Carlisle (2017), is similar to the MF in the way that it identifies undervalued companies with strong operational performance relative to their enterprise value (EV).

$$Acquirer's\ Multiple\ = \frac{Enterprise\ Value\ (EV)}{Operating\ Earnings}$$

This strategy is essentially the inverse of the EY used in the MF, and therefore this formula can be seen as a simplification. It's focused more on finding immediate "deep value" stock picks, which sometimes end up being attractive candidates for a takeover by another firm (hence the name). The top 20-30 companies with the lowest acquirer's multiple form a portfolio that follows this strategy. A potential problematic part of the formula are companies with a negative EV, which is adjusted by setting the EV to 1. It is argued that companies with large cash reserves compared to their debt and market value are exceptionally attractive investment opportunities. However, this treatment does introduce a variety of mathematical challenges, particularly for companies with negative EVs. Furthermore, this can also create unrealistic rankings, especially for companies with temporarily high cash reserves (Loughran & Wellman, 2011).

Carlisle (2017) presented evidence for its ability to outperform benchmarks by capturing 'deep-value opportunities', especially in small-cap stocks. Furthermore, Loughran and Wellman (2011) suggest that valuation metrics considering enterprise value can be effective in identifying undervalued stocks.

2.2.4 Conservative Formula

The Conservative Formula was introduced by Vliet & Koning (2017), offering a systematic approach to identifying low risk, high-return stocks. The stocks are ranked based on three key factors: volatility, price momentum, and net payout yield (NPY)[1]. The strategy focuses on selecting stocks that earn consistent returns and are comprised of companies with shareholder-friendly policies, that avoid high-risk, speculative investments. The strategy screens for the top 100 stocks with the biggest market caps that combine:

1. **Volatility:** Lowest 36-month return variance, indicative of stable performance.
2. **Price Momentum**: Stocks with upward momentum are favoured.
3. **Net Payout Yield**: Companies with high NPY are favoured.

The Conservative Formula is supported by empirical evidence. Van Vliet and Blitz (2018) showed that the strategy led to superior raw and risk-adjusted returns globally over the extensive period of 1929 to 2016. The formula achieved consistent outperformance, while maintaining lower risk than traditional equity portfolios. The study furthermore highlights the robustness across turbulent market conditions, including a lower drawdown during recessions or other instances of systemic downturns.

The static nature of the Conservative Formula limits its adaptability to evolving market conditions since it has an inherent bias towards the largest companies. It also might introduce sector biases by overweighting low-volatility industries (van Vliet & Blitz, 2018). Machine learning can address these weaknesses by dynamically adjusting the factor weights, possibly enhancing the predictive power of the CF by selecting stocks with higher momentum premiums without sacrificing the VOL or NPY rank.

2.3 Enhanced Value Metrics

Due to the utilization of value measures as a source of outperformance in all the investment formulas, further investigation into the exact metrics was warranted. Throughout the last ~15 years, there has been an increasingly convincing discussion in academia and industry, calling the value premium into question. The original components reflecting a value premium, CMA and HML, have been key pillars towards the formation of broad-market investment products, which were unfortunately seen to underperform first, after the 2000s, Loughran and Wellman (2011), Hsu (2014) and in more recent history as well Schwartz and Hanauer (2024). This multi-decade underperformance relative to its

[1]36- month historical volatility,

$$Momentum = \left(\frac{Price_{t-1}}{Price_{t-12}} - 1\right) . 100\%$$

$$NPY = \left(\frac{Common\ Div. Paid + Common\ Repurchases}{Market\ Capitalization}\right)$$

historical success spawned many debates, relating its cause to a mix of market-wide anomalies, along with a fading significance of book value as a measure. Since then, 3 alternative measures of value, along with an adjustment for operating profitability and net asset values have been proposed and regarded as being better indicators in light the overall shift in the investment world (Schwartz & Hanauer, 2024). Boudoukh et al. (2007), Walkshäusl and Lobe (2015), Walkshäusl (2016) made the initial argument that $\frac{EBITDA}{EV}$, $\frac{Operating\ CF}{Market\ Cap.}$, and NPY can be valuable additions or substitutes to the previous proxies, while Park et al. (2019), Lev and Srivastava (2019), Arnott et al. (2021), Amenc, Goltz, and Luyten (2020) argued that R&D ought to be capitalized since it is an investment in intangible assets that can provide long-term economic benefits to a company.

We utilize these propositions as part of our research by aiming to validate that NPY and $\frac{EBITDA}{EV}$ are a source of enhanced value, through the absolute and risk adjusted performance of the formulas that incorporate them. Afterward, we aim to validate the source of the returns (alpha) that they generate by regressing them onto a common factor model.

Through the addition of $\frac{Operating\ CF}{Market\ Cap.}$ in the AM, and the adjustments to net asset value and operating profitability in all formulas, we aim to add depth to out ML models. The enhanced value measure, along with the R&D adjustments to fundamental data will be incorporated alongside the existing set of features with the aims of implementing more accurate proxies within the traditional and ML formulas.

2.4 Machine Learning

Machine Learning (ML) offers a powerful way to address the traditional limitations of formula-based investing. Machine learning models can adapt dynamically to market conditions and uncover complex, non-linear relationships between financial metrics and overall stock performance (Rasekhschaffe & Jones, 2019). This makes machine learning a good fit for enhancing the aforementioned investment strategies, since the classic formulas only include stocks which score favourably on *all* of the selection criterions, with performance being tracked only in the previous 2 quarters of data.

Another advantage of ML is the possibility to incorporate additional factors traditionally overlooked. ML allows for integration of additional macroeconomic indicators, market sentiment, and other data sources. By dynamically adjusting factor weights and learning from historical patterns, ML can enhance both the accuracy and adaptability of these strategies.

To avoid the concern that the improvement by the ML method is from data mining for each strategy, rather than the sole use of machine learning as a framework, we concentrate on only one model for all classic strategies and research how it affects their cumulative performance. We considered a range of

ML models, including simple models like Logistic Regression, Decision Tree, and advanced models like Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and Neural Networks (NN). To play the may strength of ML models, that is, to capture the potential nonlinear relationships between the metrics and the stock selection, we would concentrate on advanced ML models. However, due to the limitation of our computing power, we decided to select the model from Gradient Boosting Family. Note that the specific selection of ML model is not our main research target and we expect other advanced ML models can have a similar performance as in our research.

GBM is a powerful method for handling non-linear relationships in financial data, able to combine 'weak learners' (typically decision trees), into a strong predictor (Dietterich, 2000). Within in the GBM family, we adopt XGBoost developed by Chen and Guestrin (2016), which has been proven successful in many fields. XGBoost is an advanced implementation of the gradient boosting technique. Through its built-in regularization, wise missing value handling, improved tree pruning and parallel computing, XGBoost achieves better performance and speed, and is often considered as the first choice for machine learning models containing vast amounts of data. In the field of stock selection, Rasekhschaffe and Jones (2019) found that the ML portfolio formed by XGBoost offered the competitive performance on US stocks. Gu et al. (2020) also recommend the use of gradient boosting techniques in stock selection.

Therefore, we decided to use XGBoost as the sole ML model, which would represent the potential improvement over the traditional stock selection strategies.

2.5 Gap in the Literature

Despite the empirically demonstrated historical success of traditional formula-based investment strategies, the most current research documents them as having already began to show signs of performance decay. The cumulative returns begin to stagnate even before 2019, and only partially recover in the following years before the cutoff point. While their total period performance (1963 – 2022) is well above that of the S&P 500, there is objective evidence showing signs of decline and stagnation relative to previous decades (see Appendix 1). Studies also support this notion and affirm the need for continuous innovation with the aims of keeping these approaches viable in *the long term* (Schwartz & Hanauer, 2024). While a vast body of literature behind their application exists, there is very limited exploration into their fading significance, and suggestions for their adaptation. We aim to partially fill this gap by modifying the existing formulas with enhanced value features, accounting data capitalized for R&D spending, and integration of the core framework behind each formula into XG Boost decision trees.

Furthermore, while machine learning has shown the ability to reveal non-linear and complex relationships, there is a lack of comprehensive research directly comparing traditional formulas to their

machine-learning counterparts. This leaves critical questions unanswered about the ability of machine learning to mitigate performance decay and outperform traditional models.

This study addresses these gaps by integrating machine learning techniques and traditional investment formulas. This enables the exploration of their adaptability, performance consistency, and relevance in financial markets. Specifically, it evaluates the ability of machine learning models to enhance traditional strategies allowing for better decision-making in complex, high-noise environments pertaining to today's investing climate.

3 Data Overview

The data used in this study is derived from two primary sources: the Center of Research in Security Prices (CRSP) and the CRSP/Compustat Merged Database. The CRSP dataset provides comprehensive market data for US stocks, including price, volume, and returns. The CRSP/ Compustat Merged Database gives us quarterly accounting data. These statistics are widely recognized in academic research for their robustness and reliability.

Our sample includes all US common shares, identified with CRSP codes 10 and 11. This means an exclusion of preferred stocks, ETFs, and other securities that are not common equity. We have also included delisting returns from CRSP to counteract the impact of delisted stocks on performance. To address accounting items not being provided by CRSP, we follow the methodology (Jensen et al., 2023) and Schwartz and Hanauer (2024) to calculate them manually. We also make specific adjustments to certain variables. For example, if a firm's book-to-market (BtM) value is negative or zero, we set it to missing, as is suggested by (Fama & French, 1992). This avoids distortions from accounting irregularities or extreme leverage. Furthermore, a six month lag is used to ensure that all accounting data is publicly available when the portfolio is formed, which is also suggested by (Fama & French, 1992).

The study spans from a period from 1980 to 2023, covering four decades of US equity market data. In total, the dataset includes 938,381 observations, representing a comprehensive view of firm performance and financial characteristics across market conditions. The period 1980-1999 contains 505,371 observations. The period 2000-2023 contains 444,010 observations.

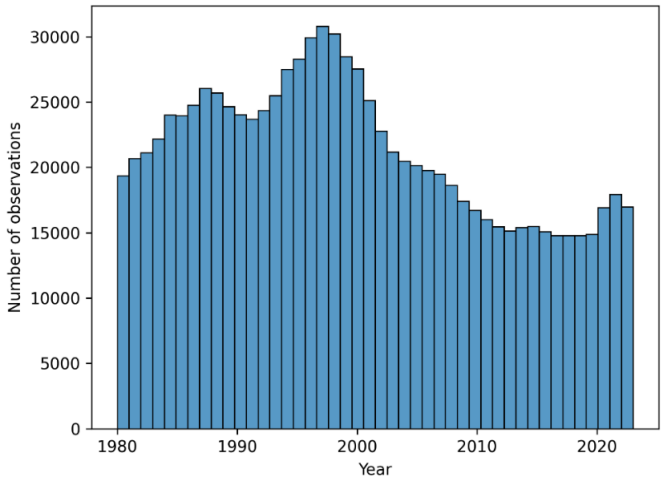


Figure 1: **The number of observations per year in the dataset.**

4 Research Methodology

We enhance the existing strategies with the use of ML models, where the traditional metrics are used as input features for the models. Some metrics may need to be transformed to better serve as input features, and additional features will be added to improve predictive accuracy as per the literature.

The analysis will compare traditional and ML enhanced strategies across time, focusing on whether ML mitigates the decay of returns, and achieves higher risk-adjusted returns. Key insights will include wether ML models beat traditional strategies overall, their efficacy in changing market regimes, assessing whether they improve as a function of time, and wether altered, or additional features serve as better indicators of outperformance.

4.1 Specification of investing formulas

To ensure the formulas adapt to the latest market dynamics, we use a more modern approach to calculate some of the metrics, which are proposed in literatures (Amenc et al., 2020; Lev & Srivastava, 2019; Schwartz & Hanauer, 2024). Formulas containing variables which involve earnings and assets (ROA, CROA, B/M, EBIT, etc.) are updated using capitalized R&D spending. This is done to reduce the stock selection bias towards firms in industries where growth is primarily driven by innovation through reinvestment (e.g. Industrials, Information Technology, Healthcare). Adjusted variables are tracked see whether they are more/less accurate and unbiased towards the prediction of security outperformance. To the Acquirer's Multiple, we add the third proposed *enhanced* value metric -Operating Cash Flow to Price ratio- where, the higher the ratio, the higher the likelihood of undervaluation. We track how, and whether each value metric becomes more/less predictive of security outperformance relative to the market index.

$$\mathrm{OCF/P} = \frac{\text{Operating CashFlow}}{\text{Market Capitalization}}$$

Based on the original papers' methodology and these updates, we recreate the following variables pertaining to each formula[2]:

F-Score: 9 binary variables split across profitability, leverage and operational efficiency dimensions: ROA, CROA, ΔROA, Accruals; ΔLEV, ΔLIQ, EQUIS; ΔGM, ΔTURNOVER

[2] All accounting metrics are gathered and implemented on a quarterly basis with a lag of 6 months (2 quarters)
All investment formulas form the portfolios based on an equal weighting of top-ranking stocks.
All formulas rebalance portfolios on a quarterly basis with updated *quarterly accounting metrics* and the latest *monthly market data*.

Selection based on the lowest decile of B/M companies, which also have an F-Score higher or equal to 7.

Magic Formula: Earnings Yield $= \frac{\text{EBIT}}{\text{Enterprise Value}}$, ROC $= \frac{\text{EBIT}}{\text{Net Fixed Assets + WC}}$

Selection based on a minimum market capitalization of $50 million, only US stocks, no foreign listings, utilities, and financial sector securities excluded. Rank EY and ROC in descending orders. The portfolio is comprised of the top 30 firms based on the sum of EY rank and ROC rank.

Acquirer's Multiple: Earnings Multiple $= \frac{\text{EV}}{\text{EBIT}}$, $\text{OCF/P} = \frac{\text{OCF}}{\text{Market Value}}$

Selection based on a minimum market capitalization of $50 million, only US stocks, no foreign listings, utilities, and financial sector securities excluded. Rank Earnings Multiple in an ascending order and OCF/P in a descending order. The portfolio is comprised of the top 30 firms based on the sum of Earnings Multiple rank and OCF/P rank.

Conservative Formula: 12-1 price MOM, 36-month average volatility, NPY

Includes foreign listings and securities from the utilities and financial sector. Rank MOM and NPY in a descending order, and VOL in an ascending order. From the 1000 largest companies, the portfolio is comprised of the top 100 stocks with the highest composite score: sum of MOM rank, Volatility rank, and NPY rank.

4.2 Evaluation methods

To proceed with evaluation, we create the baseline for comparison by back testing the 4 formulas within the historic period (2010 – 2024). Cumulative returns post-2010 are observed for levels of performance decay and fluctuations during general periods of market turmoil (2020, 2022). This comprises the baseline to which the machine learning models will be compared against.

Furthermore, we assess the efficacy of our proposed solution compared to the traditional formulas, and the role of their features throughout time. We emphasize differences between absolute and risk-adjusted performance, levels of decay overtime and overall consistency in cumulative returns. The key questions the results aim to address are whether the ML counterparts only outperform during market upswings and underperforming through crises, and if the combination of ML formulas offer a statistically significant alpha over the original formulas. We display individual graphs of each traditional formula and its ML version, and a weighted average combination of the 4. For feature adequacy we display the strength and robustness of each 3 enhanced value variables in the form heatmaps, along with line charts indicating the relevance of each formula's components, and their effectiveness towards predicting outperformance.

Finally, we assess the sources of the traditional and ML models' performance by regressing returns onto a common FF5+MOM factor model and display sources of returns.

4.3 Selection of the ML model

To highlight the ability of ML models to capture the non-linear relationships between features and the label, we prefer to adopt a model with high fitting ability with the data. However, as we decide to train a model for each quarterly prediction level, and we need to tune the hyperparameters for each model, we are sensitive to the time consumptions of training. Table 1 gives an insight of these characteristics on some candidate ML models. To balance fitting ability and training time, we select the model from Random Forest and Gradient Boosting machines. Gradient Boosting Machines have shown its higher accuracy than the Random Forest on a range of datasets (Bentéjac et al., 2021). Among the Gradient Boosting family, XGBoost shows a great balance between accuracy and time consumption (Bentéjac et al., 2021), thus we decided to use XGBoost as the model to build ML investing formulas.

Table 1: Comparison of ML models on their fitting ability and training time

This table compare the fitting ability and training time of the some commonly used ML models. In fitting ability, 'low' means that the model assuming simple or linear relationships; 'medium' means that the model can capture non-linear relationships to a certain extent but is prone to overfitting; and 'high' means that the model can capture intricate and highly non-linear relationships and can avoid overfitting with appropriate setting of hyperparameters. In training time, 'short' means that the model can be trained almost instantaneously on our dataset; 'medium' means that the model can be trained in seconds on our dataset; and 'long' means that the model takes longer training time.

Model	Fitting Ability	Training Time
Logistic Regression	Low	Short
Decision Trees	Medium	Short
Random Forest	High	Medium
Support Vector Machines	High	Long
Naive Bayes	Low	Short
Neural Networks	High	Long
Gradient Boosting Machines	High	Medium

Our selection of XGBoost is partly because of the limitation of computing power, which introduces a limitation in our research. The adoption of Support Vector Machines and Neural Networks to build ML investing formulas is still interesting and up to research.

4.4 Forming of the ML portfolios

4.4.1 Basic framework

We use the framework used by (Rasekhschaffe & Jones, 2019) to form portfolios based on ML classifiers. Consider we have a training set and a test set. Each observation in the data set represents a stock in the previous quarter, with its performance and features being used to predict its performance in this quarter. The steps are following:

1. Label the stocks' performance in the dataset as 0-1, based on their performance relative to S&P 500 in the current quarter. For each quarter, we label all stocks which outperform S&P 500 as 1 and label the rest as 0.
2. Use the data in the training set to train a ML model for predicting the probability of outperformance. That is, to predict whether the label of each quarter-stock would be 0-1. The features used as the input for the ML model per formula will be discussed in the next section.
3. Predict the probability of outperforming for each stock in a given quarter in the test set, using the trained ML model. Rank all stocks in each quarter in descending order based on these probabilities.
4. The ML portfolio for each quarter is composed of the top 10% stocks with the highest odds of outperforming, based on historical quarterly fundamental and market data.

4.4.2 Rolling training set

To ensure that the most recent market dynamics are included in the ML models, we train a new model for *each* quarter using all available historical observations. That is, the end of the training set is always the last quarter before each quarterly prediction level. For example, the training set for predicting the probabilities in Q1 2010 consists of observations from Q1 1980 to Q4 2009, while the training set for predicting probabilities in Q3 2014 consists of observations from Q1 1980 to Q2 2014.

4.4.3 Tuning of hyperparameters

To achieve better performance for models, we take the observations in the last 4 years as the validation set, where the hyperparameters of the models are tuned on it before the final training. For example, the range of training set for predicting Q1 2010 is from Q1 1980 to Q4 2009. Thus, the range of its validation set is from Q1 2006 to Q4 2009 and during the tuning process, and the range of its training set is from Q1 1980 to Q4 2005. The model with lower loss function value on the validation set is considered better.

Table 2: Candidate and initial values of hyperparameters when tuning the models.

Hyperparameter in XGBoost classifier	Candidate values	Initial value
max_depth	[3, 4, 5, 6, 7]	6

min_child_weight	[1, 3, 5]	1
gamma	[0, 1, 10, 20]	0
subsample	[0.6, 0.8, 1.0]	0.8
colsample_bytree	[0.6, 0.8, 1.0]	0.8
reg_alpha	[0, 0.1, 1, 5]	0
reg_lambda	[0, 1, 5]	0
learning_rate	[0.01, 0.05, 0.1]	0.1
n_estimators	[10, 20, 30, 50, 75, 100]	30

The candidate values of hyperparameters during the tuning are listed in Table 2. We apply a stepwise grid search when tuning the hyperparameters. That is, we categorize all the hyperparameters into four subsets. The first subset includes "max_depth", "min_child_weight" and "gamma"; the second subset includes "subsample" and "colsample_bytree"; the third subset includes "reg_alpha" and "reg_lambda"; and the fourth subset includes "learning_rate" and "n_estimators". We start from a set of initial parameters, which is also listed in Table 2, and tune these subsets of hyperparameters in sequence. A grid search is used in tuning each subset of hyperparameters. We use this approach instead of a grid search over all hyperparameters because that will lead to a gigantic number of search and thus is prone to overfitting to the validation set. The order of subsets of hyperparameters to be tuned is based on their importance in the tree model (Putatunda & Rama, 2018). The greater impact the hyperparameters have on the tree complexity, the earlier they are tuned. The candidate values include a commonly used range of values for these hyperparameters from past literatures (Bentéjac et al., 2021; Kavzoglu & Teke, 2022; Putatunda & Rama, 2018), except for the n_estimators, which we limit to relatively small numbers as the signal-to-noise ratio in financial market data is low and a high n_estimators is prone to overfitting. After the tuning, the model is finally trained on the full training set using the optimal set of hyperparameters.

4.4.4 Oversampling to emphasize recent observations.

The market changes over time, and thus recent observations are thought of as being more relevant for predicting future values. Therefore, we assign more weights to observations within recent five years through oversampling. Oversampling is a commonly used data balancing technique, which oversamples observations from the minority class to achieve a balance of data (Mohammed et al., 2020). In our case, we divide the training set into two parts, one consisting of observations within five years, and the other consisting of the rest of the observations. For example, we are training the model for predicting Q1 2010, and thus the training set consists of observations from Q1 1980 to Q4 2009. We divide it into two parts, Q1 2005 to Q4 2009 and Q1 1980 to Q4 2004. Denote the number of observations in the first part as n and that in the second part as m. In our settings, $n < m$ is always true. We oversample data in

the first part through a combination of duplication and sampling. That is, we duplicate the observations in the first part $\left\lfloor \frac{m}{n} \right\rfloor$ times, then sample $m \bmod n$ observations from the first part and combine them to form the new first part. After this oversampling, the first and second parts have same number of observations. Then we combine these parts into a new training set.

Note that this processing is after the tuning of hyperparameters and has no impact on the tuning itself.

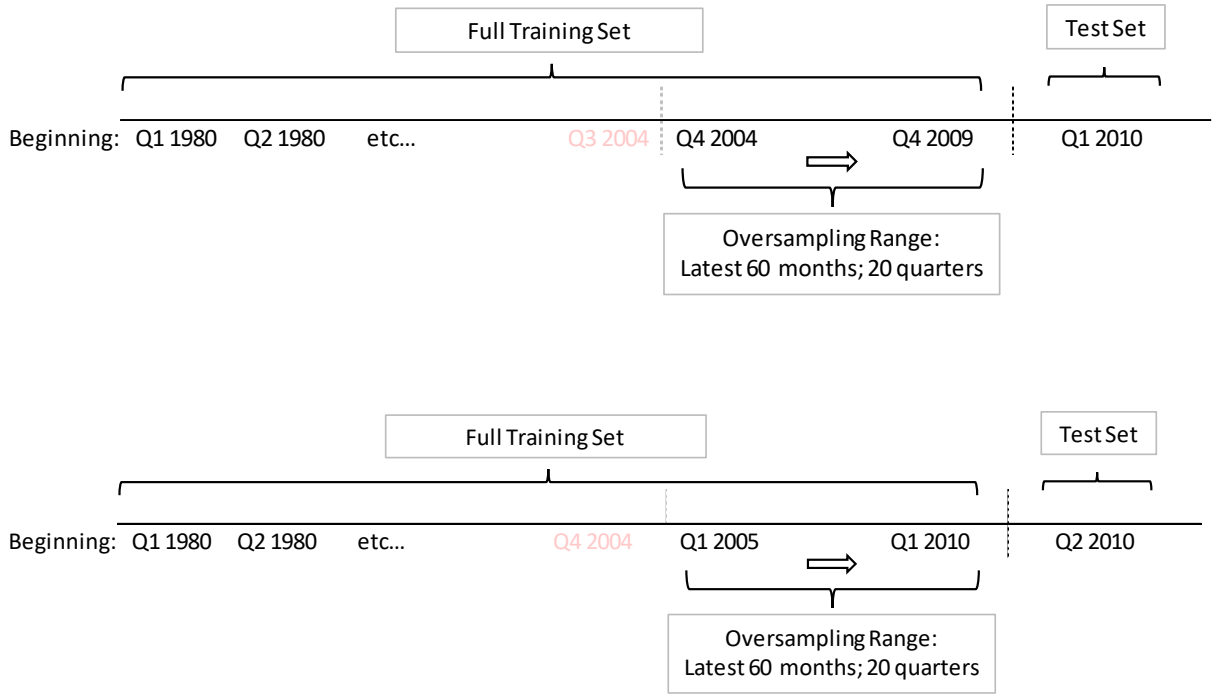


Figure 2: Visual representation of initial training and test sets

This figure illustrates the range of the full training set and for oversampling in the rolling setup. The first subplot illustrates the situation when we are training the model for predicting Q1 2010, and the second subplot illustrates the situation for Q2 2010.

4.5 Specification of the ML version strategies

In the original formulas, the ranks of the metrics are used to form the portfolios. Considering the total number of firms changes over time, we use the quantiles of the metrics instead of the ranks as the input features of the ML models.

F-Score: The features are the nine binary signals in original F-Score plus one signal being 1 if the firm's B/M is among the highest quintile, otherwise 0. This signal is named "is_cheap". Firms with a market capitalization of less than $ 50 million are excluded.

Magic Formula: The features are the quantiles of the metrics in the original Magic Formula, including the EV quantile and the ROC quantile. Firms from the financial and utility sectors and with a market capitalization of less than $50 million are excluded.

Acquirer's Multiple: The features are the quantiles of the metrics in the original Acquirer's Multiple, including the Earnings Multiple quantile and the OCF/P quantile. Firms from the financial and utility sectors and with a market capitalization of less than $50 million are excluded.

Conservative Formula: The features are the quantiles of the metrics in the original Conservative Formula, including 36-months historical volatility quantile, NPY quantile, and momentum quantile. Portfolios are formed using the largest 1,000 firms in market capitalization.

4.6 Evaluation Framework

After covering the grounds for each investment formula's execution and proposing the utilization of the ML method, the portfolio formation, rebalancing, and evaluation takes place. There is no uniform method for stock inclusion across all the formulas; the F-Score proposes the lowest BtM decile of companies with 7 or more points, the AM and MF guidelines suggest a portfolio of 20-30 stocks based on the ranking, and the CF chooses the 100 highest scoring stocks from the 1000 largest firms. In order to ensure replicability and be impartial towards each formula's variable position size, this study uses a uniform approach which is similar to others, whereby the stocks from the top decile are taken to form each of the quarterly portfolios. The decile formation is resulting from the features of each classic and ML formula. In other words, the top 10% of stocks is comprised of companies which have the best ranking, and the ranking is calculated according to the original methodology, also explained above. Companies with the highest/lowest features (highest NPY, lowest Volatility, highest EY, lowest BtM ratio, etc.) at the end of any given quarter have the highest composite ranking (combination of feature scores), hence they are they are most likely to outperform the index.

For the evaluation portion, this study adopts a long only strategy for all portfolios and all time periods as it is straightforward and poses the least amount of implementation frictions in a real-world setting. Although it is standard practice for more theoretically inclined financial studies to research investment/economic approaches using a long-short method, it is not pertinent to this application. There is ample evidence that these strategies have monotonically increasing returns from the lowest to the highest decile, with the bottom 10% providing a positive, yet significantly lower return in comparison to the top 10% (Carlisle, 2017; Greenblatt, 2006, 2010; Piotroski, 2000; Schwartz & Hanauer, 2024; van Vliet & Blitz, 2018; van Vliet & Koning, 2017). In light of this although a long-short-strategy could provide more depth and conform with standard practices, it would also be redundant within this particular context for two key reasons:

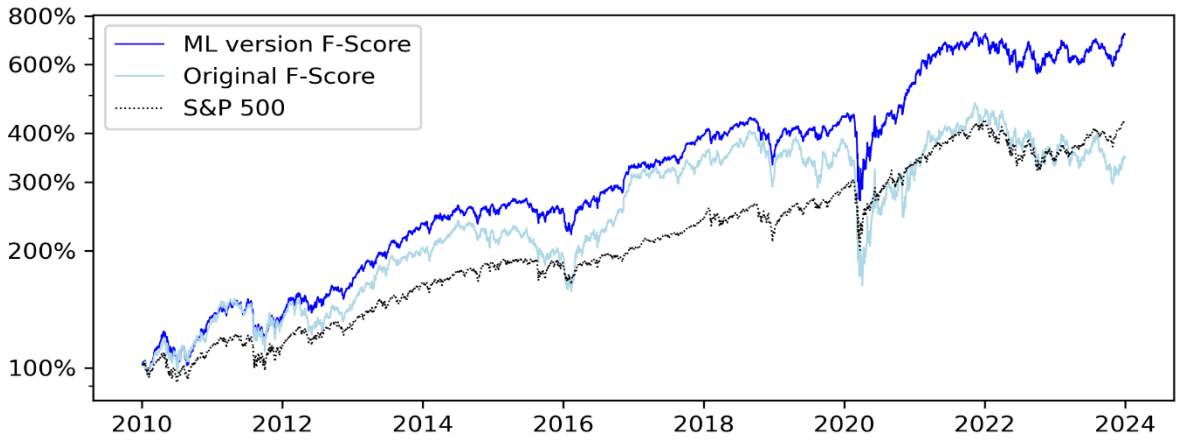
The aim is not to assess the validity of the methods and deem them as viable/unviable according to long-short return patterns, since they have already been covered and validated extensively.

A long-short portfolio would detract from the cumulative returns' magnitude, as the bottom deciles also have positive and significant returns over long periods of time. This eliminates their practical use, as an investor who would continuously implement them would only reduce their total returns, unnecessarily add complexity, and trading costs to the algorithm.

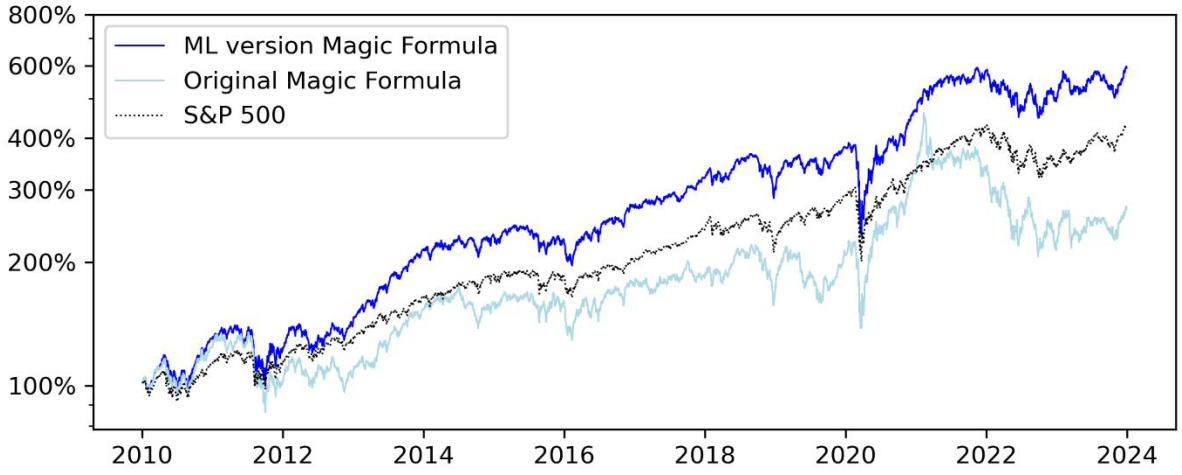
5 Results

5.1 Holistic Overview

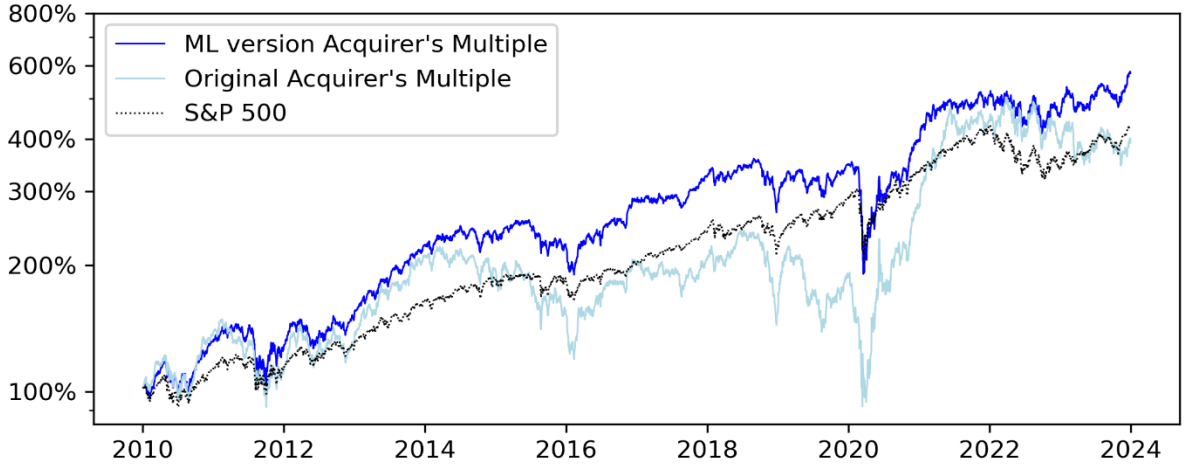
F- Score



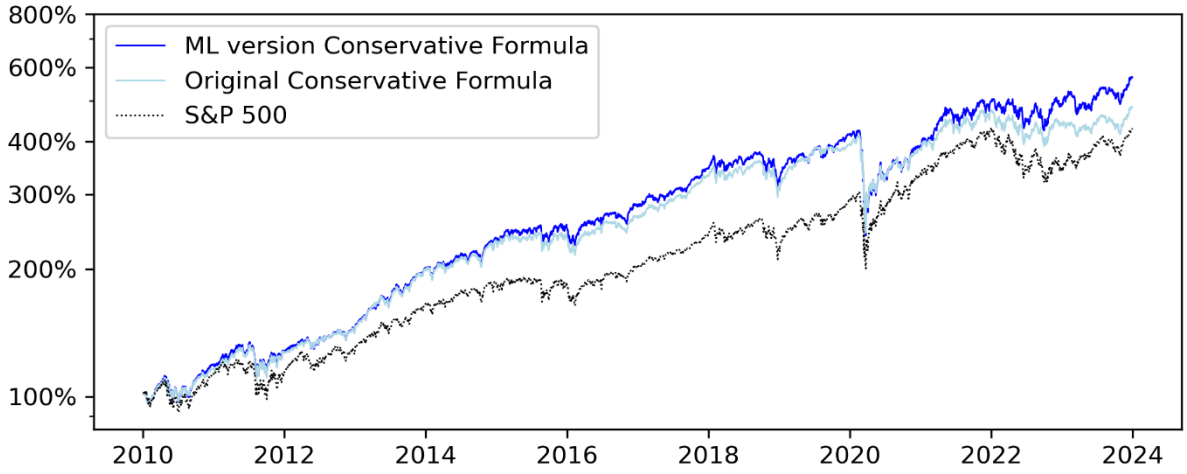
Magic Formula



Acquirer's Multiple



Conservative Formula



Combination of formulas; Equal Weighted of 4

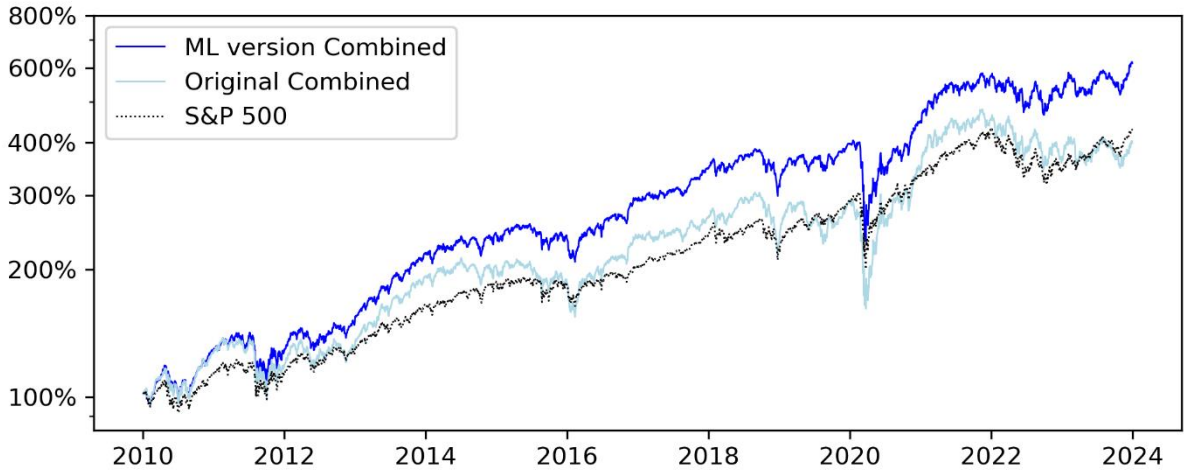


Figure 3: The cumulative performance of the formulas.

The plots represent the performance of machine learning models for each formula, with the last one being an equal weighted combination of the 4. Their cumulative returns are contrasted with those of the classic investment formulas and the S&P 500 index for the Q1 2010- Q4 2024 period.

The objective of this analysis is to compare the classic investment formulas with their ML versions in terms of absolute and risk adjusted performance, as well as display their consistency across the test period. The results are measured via long-only portfolios comprising the top decile ranking firms per formula. As opposed to decile-based long-short portfolios, this approach was taken to reflect their practical use by retail investors who face limitations in short selling and leverage.

The ML version of the F-Score emerges as the clear winner in an absolute and risk adjusted basis, while also having the smallest drawdown from each point of comparison, excluding the S&P 500. In contrast to previous studies (Carlisle, 2017; Schwartz & Hanauer, 2024), when the F-Score is combined with machine learning, it deviates from the proposition that the price at which a stock is purchased serves as the most/only important metric for future returns. While the classic Acquirer's Multiple outperforms

both formulas that include profitability measured by fundamental data, the notion of purchase price being an "be-all end-all" indicator can be seen to fade after repurposing the formulas into a ML algorithm.

Thus, in the context of an ML enhancement, this serves as evidence towards the arguments that profitability is an equally, *if not more*, important aspect leading to outperformance. Furthermore, the form and number of measures used to assess profitability can be seen to add increasingly valuable information, extending beyond the 3 features used in the Magic Formula.

As for the Conservative formula, it had the smallest improvement over the classic iteration. This suggests that it is the least prone to having non-linear/complex relationships which can be implicitly uncovered by ML. The exact cause of the tighter gap between classic and ML, along with it having the *highest classic* cumulative return out of all formulas is open to interpretation, although there are two pre-existing claims that could help explain it:

The onset and increase in performance decay- the book and study documenting Conservative formula investing were published in 2017 and 2018, while the F-Score, Magic Formula and Acquirer's Multiple preceded it by 17, 11 and 5 years respectively, leaving much less time for markets to continuously implement and exhaust the edge it offers.

The ever-increasing gap between high and low B/M companies leading to pervasive and elevated valuation spreads, causing traditional "value" strategies to underperform considerably in comparison to "growth" within the previous decades. As the CF only relies on market data, and selects from the largest 1000 firms, it doesn't screen for value explicitly and also has the lowest SMB coefficient from all formulas in the extended 5 factor model (0.08 and 0.16 for classic and ML).

Overall, the individual and combined performance of the classic formulas within the test period is reaches parity with the S&P 500 and even begin trailing it within the last ~2 years. They also tended to underperform considerably during general downturns. In contrast to previous papers which had a longer historical testing period, the timeframe and current results represent a tipping point for 3 out of 4 approaches, rendering them increasingly inferior. The evidence is substantial and further validates the need for continuous innovation, which was provided via feature enhancement and the implementation of machine learning in this study.

5.2 Detailed assessment per formula

Table 3: Performance Criteria.

This table reports the cumulative returns, compounded annual average return, standard deviation, Sharpe ratio and maximal drawdown for each portfolio according to its strategy, along with the S&P 500 index serving as the baseline for comparison.

Q1 2010 – Q4 2023	Cumulative	CAGR	St. Dev.	Sharpe Ratio	M.D.D.
Original F-Score	344.6%	9.24%	26.4%	0.44	(59.9%)
ML F-Score	714.7%	15.08%	20.1%	0.76	(40.4%)
Original Magic Formula	269.0%	7.32%	24.2%	0.38	(55.2%)
ML Magic Formula	592.8%	13.56%	20.1%	0.69	(40.2%)
Original Acquirer's Multiple	391.4%	10.24%	29.4%	0.45	(62%)
ML Acquirer's Multiple	574.1%	13.30%	21.4%	0.65	(47%)
Original Conservative Formula	482.3%	11.90%	16.5%	0.71	(41%)
ML Conservative Formula	567.0%	13.20%	17.6%	0.74	(43.7%)
Original Combined	398.5%	10.38%	21.7%	0.52	(47%)
ML Combined	615.9%	13.87%	19.5%	0.72	(42.3%)
S&P 500	427.7%	10.94%	17.4%	0.63	(33.9%)

Table 4: Regression results on the factor model.

This table reports the regression results of all original formulas and ML formulas on the Fama-French five-factor model plus momentum. Daily returns of these formulas from 2010 to 2023 are regressed on Fama-French five factors plus momentum. Column α reports the constants in the regressions in an annualized frequency. Column RMRF, SMB, HML, RMW, CMA and MOM report the loadings of these formulas on the factors. Newey–West t-statistic (Newey & West, 1987) for each coefficient is reported in parentheses below it. One-sided significance tests for α are performed. Significance levels at 90%, 95%, and 99% are denoted by one, two, and three stars, respectively.

	α	RMRF	SMB	HML	RMW	CMA	MOM
Original F-Score	0.59%	0.96	0.74	0.49	-0.01	-0.11	-0.29
	(0.18)	(40.78)	(22.69)	(12.99)	(-0.16)	(-1.89)	(-11.12)
ML F-Score	3.35%	0.94	0.62	0.20	0.05	-0.04	-0.03
	(3.93***)	(123.82)	(47.65)	(15.68)	(4.31)	(-1.98)	(-3.10)
Original Magic Formula	-2.01%	0.93	0.83	-0.18	-0.24	0.06	-0.13
	(-0.64)	(58.62)	(30.79)	(-6.17)	(-6.84)	(1.08)	(-5.87)
ML Magic Formula	0.99%	0.98	0.62	0.03	0.15	0.04	-0.03
	(1.00)	(151.22)	(47.48)	(1.96)	(12.28)	(1.69)	(-3.40)
Original Acquirer's Multiple	0.95%	1.03	0.89	0.34	0.04	0.07	-0.31
	(0.23)	(17.43)	(10.67)	(5.97)	(0.60)	(0.86)	(-8.65)
ML Acquirer's Multiple	0.87%	1.00	0.70	0.17	0.20	0.03	-0.08
	(0.81)	(139.66)	(51.62)	(12.19)	(14.05)	(1.16)	(-6.93)
Original Conservative Formula	-0.74%	0.89	0.08	0.19	0.22	0.09	0.19
	(-0.58)	(64.82)	(4.46)	(9.84)	(13.42)	(3.26)	(17.36)
ML Conservative Formula	0.51%	0.93	0.16	0.30	0.25	-0.01	0.11
	(0.43)	(76.54)	(9.15)	(16.20)	(15.89)	(-0.46)	(11.67)
Original Combined	-0.30%	0.95	0.63	0.21	0.00	0.02	-0.13
	(-0.16)	(45.69)	(23.54)	(8.66)	(0.11)	(0.75)	(-7.46)
ML Combined	1.43%	0.96	0.52	0.17	0.16	0.00	-0.01
	(1.92**)	(135.98)	(42.93)	(14.20)	(17.02)	(0.14)	(-0.62)

After covering the return consistency and differentials within the sample period, we measure the extent to which each of the strategies' performance is attributable to common factors. The long only portfolios' returns are regressed onto the factors in the FF5 model, with the extension of momentum.

All ML iterations generate positive alpha over the classic approaches, except the Acquirer's multiple, and all their alpha are substantially greater than 0, although only the ML version of the F-Score and the equally weighted ML combination are statistically significant at the 1% and 5% levels respectively. All ML iterations except the ML Conservative formula offer a considerably lower max-drawdown which is solely the result of the XGBoost framework, which allowed for dynamic feature selection, also resulting in better risk-adjusted performance. All ML and classic formulas had the same set of features and adjusted accounting data.

None of the formulas overweigh the market factor even though they all have a notably larger max drawdown relative to the S&P500 for the sample period. There is a marked decrease in SMB and, to an extent, HML for the ML F-Score, MF and AM formulas, while the CF remains low in both examples. The CF biases towards large firms, as they tend to be less volatile and have higher payout yields due to their limited growth opportunities (Bulan & Yan, 2010). By the same breadth, the moderately negative tilt in momentum neutralizes from classic to ML, except for the CF, which is expected since it has an inbuilt momentum preference. Across all formulas, this slight decrease in SMB, HML and increase in MOM coefficients leads to a relative cumulative performance addition of ~200% for this 14-year period. This confirms that these 3 classic strategies' edge *in the past* largely arose from an intelligent way of capturing a mix of value and quality stocks with terrible price performance in recent history (negative MOM), which also drove returns during times when small caps were favoured. The relative increase in MOM coefficient from classic to ML also supports the fact that, all else equal, a positive or neutral exposure to momentum would increase returns of these strategies as the premium is large and significant over time.

All formulas except the classic MF exhibit positive exposure to value (HML), with the AM and the F-Score having the highest loadings. The AM focused solely on cheap companies via the enterprise multiple and the OCF-to-price, and the F-Score directly incorporates the B/M ratio for portfolio selection. For the Magic Formula, the HML exposure is negative, and when transformed into ML it neutralizes and adds a moderate tilt towards RMW, unlike any other formula. Since the fundamental data incorporated capitalized for R&D (which affected profitability and asset levels), but the traditional factor model premiums for HML and RMW *by default do not*, this can translate into a moderate skew in the regression results for all formulas. R&D increases net fixed assets and EBITDA and decreases EBIT. Since the MF relies only on two features, both with altered fundamental data, it leads to lower earnings yields (EY) and lower returns on capital (ROC), resulting in a worse composite rank. Due to this, the Magic formula favours companies that *appear* as smaller B/M-wise and have weak(er)

operating profitability in a regression. All of this is corrected and accounted for within the ML MF, as it's *outperformance* is statistically significant at the 5% level, over the classic.

For the Conservative Formula, the positive value exposure can be attributed to the selection criteria from the NPY component. Results are consistent with previous analyses (van Vliet & Blitz, 2018), reporting a positive HML exposure in the subperiod from 1963 to 2016.

Table 5: The t-test results for difference between risk-adjusted returns.

This table reports the t test results for the difference of risk-adjusted daily returns between the ML and original versions of each formula. The risk-adjusted return is computed as $R_{adj,t} = \frac{R_t - r_{f,t}}{\sigma(R_t)}$. Significance levels at 90%, 95%, and 99% are denoted by one, two, and three stars, respectively.

	F-Score	Magic Formula	Acquirer's Multiple	Conservative Formula	Combined
t-statistic	2.29**	2.21**	1.31	0.57	2.19**

As an addition to the performance analysis, the incremental gain in terms of risk adjusted returns from classic to ML was assessed. The ML versions of F-Score, Magic Formula and overall combination of the formulas present a significant improvement in returns quality at the 5% level. In absolute terms the proposed solution resulted in a gain of 217.4% over the base formulas, and 188.6% over the index from Q1 2010 – Q4 2023. The combination of the original strategies can be seen to slightly underperform the index by 29.2%, with the biggest detraction coming from the Magic Formula.

While these returns are reported on a gross basis, adjustments for trading costs would have a moderate impact on the portfolios, however the significance of this performance would remain similar, as both the classic and ML strategies employ the same rebalancing and stock selection principles on a quarterly basis. Across the entire duration, on average, the F-Score portfolios held 333 stocks, the MF and AM 266 and the CF had 100. These were the de-facto top deciles for each formula. Trading costs are heavily dependent on the dollar amount invested in each position and the broker/plans offered, therefore there is no straightforward answer.

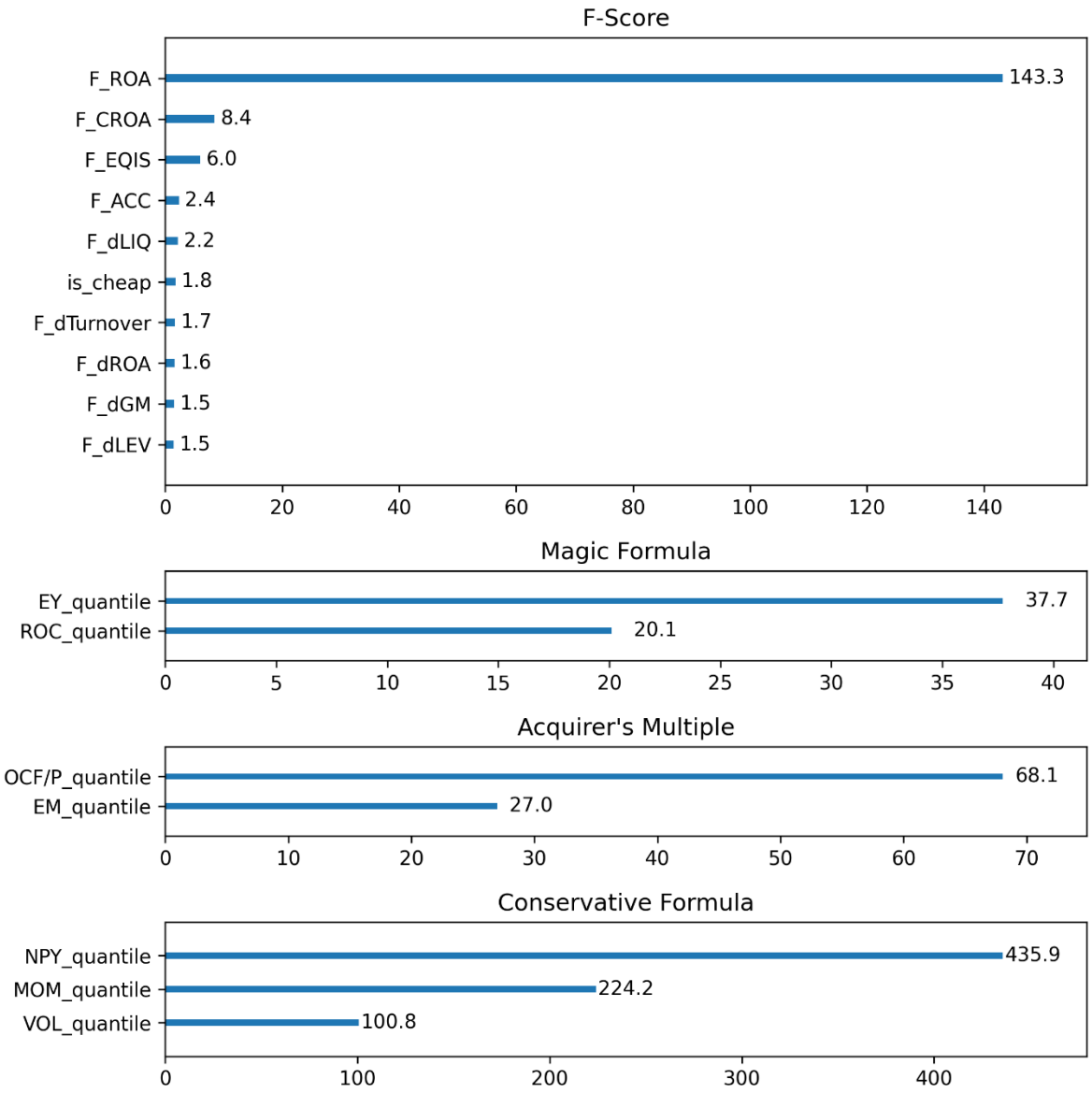


Figure 4: Importance of features in ML formulas.

This figure plots the importance of features by the average gain of splits which use the feature. A split refers to the division of data at a tree node into two branches based on a threshold value of a feature, chosen to minimize the model's loss function. In the fitting process of the XGBoost model, after adding a split node, the model's loss function value decreases, and this decrease is called 'gain'. The average gain of splits using a feature is computed as the feature's importance in the model.

In examining the feature importance from the machine learning models, we assess their contribution towards predicting the probabilities of outperforming the index.

In this application, splitting stocks based on their ROA, Earnings Yield, OCF-to-Price, and Net Payout Yield scores, offers the greatest gain for that model's overall predictive capabilities. In more straightforward terms, in the Conservative Formula for example, distinguishing between stocks based

on low and high NPY quantiles offers a more robust estimate that would indicate future outperformance than "12-1" momentum or 36-month volatility.

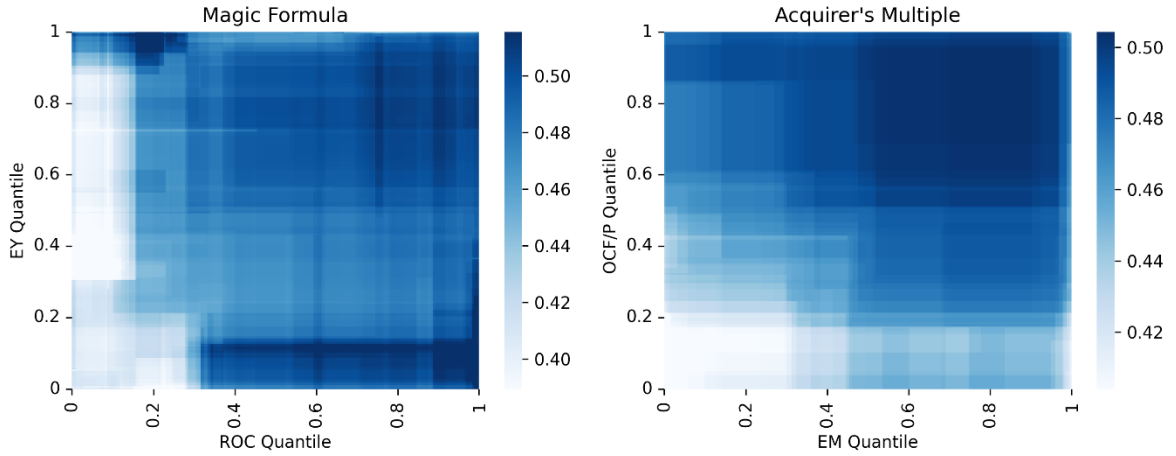


Figure 5: Models predicted probabilities heatmap.

This figure plots the predicted probabilities of the ML Magic Formula model and the ML Acquirer's Multiple model. Each point in the left subplot represents the probability of outperforming predicted by the ML Magic Formula Model based on its corresponding EY and ROC quantiles. Each point in the right subplot represents the probability of outperforming predicted by the ML Acquirer's Model based on its corresponding EM and OCF/P quantiles. EM quantile is in a decreasing arrangement, while others are in increasing arrangements. The color bar per subplot represents the corresponding color for different probabilities of outperforming.

For the formulas containing only 2 of the enhanced value features, the efficacy and interdependence leading to the models' classification is shown through the 2-dimensional heatmaps.

In the ML Acquirer's Multiple, there is a linearly increasing pattern indicating that when all else are equal, the $\text{Earnings Multiple} = \frac{\text{EV}}{\text{EBIT}}$ and $\text{OCF/P} = \frac{\text{OCF}}{\text{Market Value}}$ higher, the stronger the chances that stock will outperform. This is expected, since both formulas assess profitability levels with regards to firm size. Nevertheless, alluding to the feature importance analysis covered previously, the cash-flow measure provided a higher gain rate, suggesting that firms which are strictly profitable on a cash basis in relation to their market capitalization have higher odds of beating the index. Furthermore, OCF/P can be perceived as being a more immediate proxy for profit generation in monetary terms since it is less subjected to non-cash accounting measures, or other accounting adjustment practices native to GAAP.

In the ML Magic Formula, the dependency of these features is not entirely aligned and straightforward. The most saturated areas on the heatmap indicate that either, firms with high $\text{Earnings Yield} =$

$\frac{\text{EBIT}}{\text{Enterprise Value}}$ and *low* $\text{ROC} = \frac{\text{EBIT}}{\text{Net Fixed Assets + WC}}$, or firms with the high ROC and *low* Earnings Yield are more likely to outperform.

While the traditional Magic Formula emphasizes a strong composite rank (High EY + High ROC), this analysis shows that the use of the EY metric in the top quantiles could be a sufficient and viable indicator, provided that ROC is kept at the bottom quantiles, and vice versa. Furthermore, only when both metrics are taken to their extremes to form a high composite score, then they also exhibit a similar, though lower overall probability.

The different structure of relationships between the features and the outputs in ML Acquirer's Multiple and in ML Magic Formula can explain why the ML Magic Formula significantly beat original Magic Formula, but the ML Acquirer's Multiple doesn't. When the relationship is quite monotonous, a ML approach cannot substantially beat a rank approach by a lot, as the rank approach can also easily find the best stocks. However, when the relationship is non-monotonous, the rank approach may not find the best stocks, while the ML approach still works through finding the non-monotonous pattern of relationships in historical data. The performance of ML Magic Formula shows the ability of ML approach to capture the non-linear relationships in the metrics and help stock-selection.

6 Discussion

The results of this study provide evidence for the value of integration of machine learning techniques into traditional investment formulas. By modernizing these frameworks, we address performance decay and explore how machine learning models could enhance the adaptability and predictive power of investment strategies. In this section, we will discuss the implications of all these findings, highlight their contributions, and suggest topics for future research.

6.1 Performance of ML-Enhanced Strategies

The study highlights the consistent outperformance of the machine learning-enhanced strategies compared to their traditional counterparts. This shows the adaptability and flexibility that machine learning can bring to investing. For instance, the ML-enhanced F-score shows significant improvements in the cumulative returns as well as in the risk-adjusted performance, with the Sharpe Ratio increasing from 0.44 to 0.76 for the F-Score. This suggests that the ML's ability to improve the traditional formulas provides a key advantage over static, rule-based models.

Interestingly, the Conservative Formula showed the least amount of improvement compared to the other strategies. This may indicate that the relationship between the metrics in the Conservative Formula and the quality of stocks is quite monotonous, where a rank approach is enough. On the other hand, strategies that heavily rely on valuation and profitability metrics, such as the F-Score, the Magic Formula, and the Acquirer's Multiple, benefited significantly from the machine learning enhancements. This might suggest that machine learning is particularly valuable when applied to strategies where traditional metrics fail to account for time-varying relationships.

6.2 Sources of Returns

Because of the integration of additional features, the machine learning models were able to reflect the importance of non-traditional value drivers for financial analysis. This is a trend that can also be observed increasingly in accounting standards.

Another notable finding is the enhanced exposure of machine learning-driven strategies to momentum and profitability factors. For example, the machine learning-enhanced Magic Formula demonstrated a positive tilt towards these factors, which contrasts with the traditional version's weaker or negative exposure to momentum. This ability to adapt to changing drivers of performance suggests that machine learning can complement traditional frameworks by incorporating additional dimensions of stock evaluation that static models might overlook.

6.3 Implications for Systematic Investing

The findings of this research show important implications for systematic investing and for systematic investors. First, it shows that machine learning is a critical tool to use to maintain competitive and efficient in the current financial market. The performance decline of static strategies furthermore highlights the importance of dynamic methods that allow for quick response to changing market conditions.

For investors, the benefits of using machine learning are large. The enhanced strategies offer significant advantages in high-noise and competitive environments. The findings suggest that practitioners who integrate machine learning into their investment process are better equipped to construct portfolios that achieve superior risk-adjusted returns.

6.4 Limitations and Future Research

While the study shows promising results, there are also some limitations which should be taken into account. First, the study is limited to US equity only. This leads to questions regarding the geographic generalizability of the results. Future research might focus on extending the framework to be generalizable across locations.

Furthermore, we have used a single machine learning model for all the strategies due to its ease of use and speed of execution. While this is a strong model, the inclusion of different models might also be interesting for future research. Multiple machine learning models can also be combined to try to enhance the strategies even further.

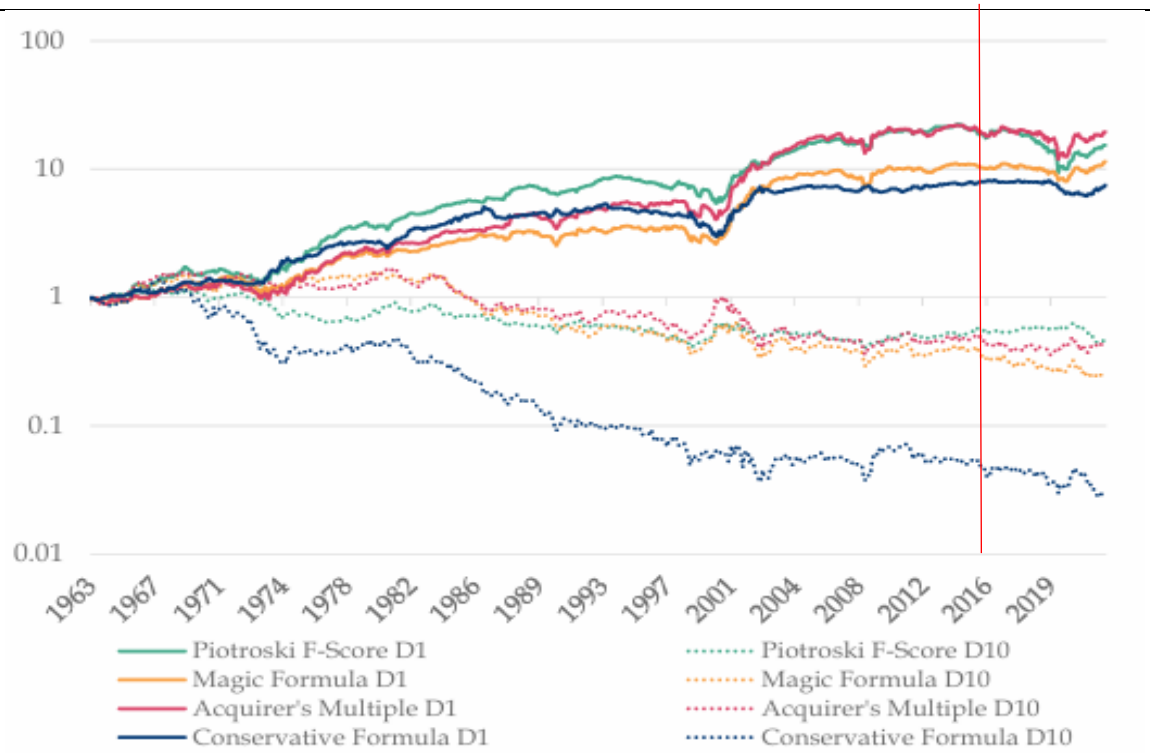
Lastly, we have focused on long-only portfolios and left out practical factors such as trading costs, partial or full illiquidity frictions and taxation. The study design had the retail investor as the prime beneficiary of such strategies, and the formulas themselves originally had the same target audience in mind as well. Future research is needed in order to determine the scaling effects of trading costs that apply the rates from professional brokerage accounts and also measure the liquidity spread for each transaction, along with a more in depth analysis featuring a short side portfolio.

6.5 Conclusion

The integration of machine learning into traditional investment strategies represents a significant advancement in systematic investing. Machine learning enhances adaptability, accuracy, and performance. The findings contribute to both academic literature and practical investing, offering a clear path for the modernization of the modernization of formula-based investing. As financial markets keep evolving, machine learning will provide a powerful tool for achieving sustainable outperformance.

7 Appendices

Appendix 1



Year	F- Score D1	MF D1	AM D1	CF D1	Market
2000	24%	20%	23%	23%	-9%
2001	20%	66%	50%	15%	-8%
2002	-3%	2%	-10%	4%	-23%
2003	45%	48%	50%	16%	26%
2004	18%	32%	28%	15%	10%
2005	18%	4%	14%	12%	6%
2006	22%	18%	18%	13%	13%
2007	6%	3%	-3%	4%	12%
2008	-36%	-41%	-36%	-30%	-35%
2009	65%	56%	71%	13%	28%
2010	18%	19%	23%	20%	17%
2011	9%	8%	2%	12%	6%
2012	28%	9%	10%	13%	14%
2013	33%	48%	39%	32%	28%
2014	11%	9%	4%	15%	12%
2015	-10%	-5%	-15%	10%	1%
2016	30%	19%	35%	11%	12%
2017	*13%*	*13%*	*13%*	*23%*	*20%*
2018	*-25%*	*-9%*	*-16%*	*-4%*	*-2 %*
2019	*13%*	*29%*	*26%*	*22%*	*29%*
2020	*20%*	*24%*	*30%*	*8%*	*27%*
2021	*33%*	*27%*	*32%*	*20%*	*22%*
2022	*-3%*	*0%*	*2%*	*0%*	*-19%*
Total Average	15%	17%	17%	12%	8%
Last 5 Years	*9%*	*14%*	*15%*	*12%*	*13%*

Appendix 1: Taken from “Formula Investing” (Schwartz & Hanauer, 2024)

This graph displays the cumulative performance, alongside the per year simple returns of the top deciles per formula. While in earlier periods (2000-2016) the classic iterations of the formulas can be seen to greatly outperform the market, offering superior positive returns or less negative ones even when the entire market is down, there is a *notable* break in the trend post 2017. The cumulative outperformance for the 22-year period stems greatly from the first 14 years of this benchmark (and even before), while the notable trend change emerges in 2017 when all formulas except for the CF underperform the market for the first time and continue doing so until 2022. The 5-year cumulative performance still appears to be commendable, however, none of the formulas would have offered a significantly greater end result over a low-cost ETF on a gross returns basis, let alone after adjusting for trading and implementation costs.

8 References

Amenc, N., Goltz, F., & Luyten, B. (2020). Intangible capital and the value factor: Has your value definition just expired? *The Journal of Portfolio Management*.

Arnott, R. D., Harvey, C. R., Kalesnik, V., & Linnainmaa, J. T. (2021). Reports of Value's Death May Be Greatly Exaggerated. *Financial Analysts Journal*, *77*(1), 44–67. https://doi.org/10.1080/0015198X.2020.1842704

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Boudoukh, J., Michaely, R., Richardson, M., & Roberts, M. R. (2007). On the Importance of Measuring Payout Yield: Implications for Empirical Asset Pricing. *The Journal of Finance*, *62*(2), 877–915. https://doi.org/10.1111/j.1540-6261.2007.01226.x

Bulan, L. T., & Yan, Z. (2010). *Firm Maturity and the Pecking Order Theory* (SSRN Scholarly Paper 1760505). Social Science Research Network. https://doi.org/10.2139/ssrn.1760505

Carlisle, T. E. (2017). *The acquirer's multiple: How the billionaire contrarians of deep value beat the market*. Ballymore Publishing.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Deng, X. (2016). *Piotroski's F-Score in the Chinese A-Share market*. http://hdl.handle.net/11427/24520

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Multiple Classifier Systems*, 1–15. https://doi.org/10.1007/3-540-45014-9_1

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, *47*(2), 427–465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

Greenblatt, J. (2006). *The little book that beats the market*.

Greenblatt, J. (2010). *The Little Book That Still Beats the Market*. John Wiley & Sons.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Gunnar Juliao de Paula, A. (2016). *Backtesting The Magic Formula In The Brazilian Stock Market*.

Hsu, J. C. (2014). *Value Investing: Smart Beta vs. Style Indices* (SSRN Scholarly Paper 2477293). Social Science Research Network. https://papers.ssrn.com/abstract=2477293

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is There a Replication Crisis in Finance? *The Journal of Finance*, *78*(5), 2465–2518. https://doi.org/10.1111/jofi.13249

Kavzoglu, T., & Teke, A. (2022). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). *Bulletin of Engineering Geology and the Environment*, *81*(5), 201. https://doi.org/10.1007/s10064-022-02708-w

Lev, B., & Srivastava, A. (2019). *Explaining the Recent Failure of Value Investing* (SSRN Scholarly Paper 3442539). Social Science Research Network. https://doi.org/10.2139/ssrn.3442539

Loughran, T., & Wellman, J. W. (2011). New Evidence on the Relation between the Enterprise Multiple and Average Stock Returns. *Journal of Financial and Quantitative Analysis*, *46*(6), 1629–1650. https://doi.org/10.1017/S0022109011000445

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Newey, W. K., & West, K. D. (1987). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, *28*(3), 777–787. https://doi.org/10.2307/2526578

Park, H. & others. (2019). An intangible-adjusted book-to-market ratio still predicts stock returns. *Critical Finance Review*, *25*(1), 207–236.

Piotroski, J. D. (2000). Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers. *Journal of Accounting Research*, *38*, 1–41. https://doi.org/10.2307/2672906

Putatunda, S., & Rama, K. (2018). A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, 6–10. https://doi.org/10.1145/3297067.3297080

Rasekhschaffe, K. C., & Jones, R. C. (2019). Machine Learning for Stock Selection. *Financial Analysts Journal*, *75*(3), 70–88. https://doi.org/10.1080/0015198X.2019.1596678

Schwartz, M., & Hanauer, M. X. (2024). *Formula Investing* (SSRN Scholarly Paper 5043197). Social Science Research Network. https://doi.org/10.2139/ssrn.5043197

Tikkanen, J., & Äijö, J. (2018). Does the F-score improve the performance of different value investment strategies in Europe? *Journal of Asset Management*, *19*(7), 495–506. https://doi.org/10.1057/s41260-018-0098-3

van Vliet, P., & Blitz, D. (2018). *The Conservative Formula: Quantitative Investing Made Easy* (SSRN Scholarly Paper 3145152). Social Science Research Network. https://doi.org/10.2139/ssrn.3145152

van Vliet, P., & Koning, J. de. (2017). *High Returns from Low Risk: A Remarkable Stock Market Paradox*. John Wiley & Sons.

Walkshäusl, C. (2016). Net payout yields and the cross-section of international stock returns. *Journal of Asset Management*, *17*(1), 57–67. https://doi.org/10.1057/jam.2015.34

Walkshäusl, C. (2020). Piotroski's FSCORE: International evidence. *Journal of Asset Management*, *21*(2), 106–118. https://doi.org/10.1057/s41260-020-00157-2

Walkshäusl, C., & Lobe, S. (2015). The Enterprise Multiple Investment Strategy: International Evidence. *Journal of Financial and Quantitative Analysis*, *50*(4), 781–800. https://doi.org/10.1017/S002210901500023X