Machine learning and manager selection: evidence from South Africa

Daniel Page, Yudhvir Seetharam and Christo Auret
School of Economics and Finance, University of the Witwatersrand, Johannesburg, South Africa

Received 24 June 2022
Revised 10 November 2022
20 April 2023
Accepted 22 June 2023

Abstract

Purpose – This study investigates whether the skilled minority of active equity managers in emerging markets can be identified using a machine learning (ML) framework that incorporates a large set of performance characteristics.

Design/methodology/approach – The study uses a cross-section of South African active equity managers from January 2002 to December 2021. The performance characteristics are analysed using ML models, with a particular focus on gradient boosters, and naïve selection techniques such as momentum and style alpha. The out-of-sample nominal, excess and risk-adjusted returns are evaluated, and precision tests are conducted to assess the accuracy of the performance predictions.

Findings – A minority of active managers exhibit skill that results in generating alpha, even after accounting for fees, and show that ML models, particularly gradient boosters, are superior at identifying non-linearities. LightGBM (LG) achieves the highest out-of-sample nominal, excess and risk-adjusted return and proves to be the most accurate predictor of performance in precision tests. Naïve selection techniques, such as momentum and style alpha, outperform most ML models in forecasting emerging market active manager performance.

Originality/value – The authors contribute to the literature by demonstrating that a ML approach that incorporates a large set of performance characteristics can be used to identify skilled active equity managers in emerging markets. The findings suggest that both ML models and naïve selection techniques can be used to predict performance, but the former is more accurate in predicting *ex ante* performance. This study has practical implications for investment practitioners and academics interested in active asset manager performance in emerging markets.

Keywords Active asset management, Machine learning, Emerging markets, Manager skill

Paper type Research paper

1. Introduction

The performance of active asset managers has been the subject of extensive research over the past 50 years. The literature consistently documents aggregate underperformance by active managers, particularly on a net-of-fees basis. This has led to a shift towards passive investing, with United States (US) passive assets under management surpassing actively managed funds in August 2019 (see Gittelsohn, 2019 and Standard and Poor's SPIVA U.S. scorecard, 2021). Despite this trend, some studies have identified subsets of active managers across various datasets and time-periods that consistently produce positive net-of-fees alphas (see Kacperczyk *et al.*, 2005; Fama and French, 2010; Kacperczyk *et al.*, 2014; Berk and Van Binsbergen, 2015). A crucial question that arises from the literature is whether the minority of "skilled" active managers can be identified on an *ex ante* basis. Manager selection using historical data in the form of characteristics (independent variables or "features") presents an empirical question that fits neatly within a machine learning (ML) experiment. Recent literature by DeMiguel *et al.* (2021) and Kaniel *et al.* (2022) provides evidence of the predictive ability of ML techniques in generating investment alpha on the cross-section of US domestic

The authors would like to thank the Editor of the International Journal of Emerging Markets and anonymous reviewers for their helpful comments in improving their article.

International Journal of Emerging Markets
Vol. 20 No. 5, 2025
pp. 1819-1848

1746-8809
DOI 10.1108/IJOEM-06-2022-0998

equity mutual funds. This study applies several ML models to the cross-section of South African active equity managers over the period January 2002 to December 2021. To the best of our knowledge, this study is the first to consider performance predictability on a cross-section of South African active equity managers using either parametric linear or ML methods. More importantly, this study evaluates whether ML for manager selection is as robust when applied on a smaller cross-section of managers that invest in a considerably smaller, less liquid equity market.

Similar to DeMiguel *et al.* (2021) and Kaniel *et al.* (2022), we use ML techniques to predict the out-of-sample performance of active equity managers in the South African market using 36 performance specific fund characteristics measured across managers over rolling 60-month windows. The managers are then ranked and sorted into decile portfolios based on predicted performance, and out-of-sample performance is assessed. The study provides novel findings related to the application of ML and manager selection in emerging markets with smaller market bourses and narrower active manager universes.

We find that gradient boosting models, specifically LightGBM (LG) and XGBoost (XG), are superior and produce zero-cost long-short excess returns of 3.3 and 1% per annum, respectively, consistent with the literature on developed markets. We also assess the robustness of the ML results against naïve selection techniques, such as momentum, reversal and style alpha, and find that naïve ranking performs as well as most of the ML models applied but is still inferior to LG.

We further evaluate the predictive accuracy of the respective ML algorithms by considering manager selection precision within the top decile by evaluating the correlation between *ex ante* predicted and ex-post performance rank. We find evidence of the superior predictive ability of non-linear ML models in determining both future top and bottom decile active managers, confirming the literature on developed markets.

Overall, the study concludes that ML can be used to predict the future performance of active equity managers in emerging markets, and the predictive ability is exploitable. Gradient boosting and ensemble algorithms are superior due to their ability to capture non-linear interactions and relationships between features and labels, and the results are novel as they are the first conducted on a cross-section of emerging market active managers.

Importantly, this study is agnostic in-terms of the ML models applied. Therefore, there is no intention to define an optimal set of features or prescribe a ML model best suited to predict manager return for the cross-section of South African active equity managers. Rather, the essence of the research objective can be delineated into three core hypotheses. First, assess whether ML can be used to identify top and bottom performing managers on an *ex ante* basis. Second, determine if portfolios sorted on in-sample prediction results in out-of-sample nominal and risk-adjusted profits; and third, determine whether findings for a smaller universe of emerging market active equity managers are consistent and/or comparable with developed market literature.

The rest of the paper is structured as follows. Section 2 provides a brief description of the pertinent literature; Section 3 details the research design, data and methodology; Section 4 provides the results, and Section 5 concludes.

2. Literature review

Active manager underperformance is a well-documented phenomenon in the international literature. Various studies have examined fund-specific characteristics that can explain underperformance, including fund size (Chen *et al.*, 2004), fees (Gil-Bazo and Ruiz-Verdú, 2009), diseconomies of scale (Berk and Green, 2004), age (Blake and Morey, 2000) and fund manager attributes such as education and experience (Gallagher, 2003; Brands *et al.*, 2005), as well as recent historical performance (Grinblatt and Titman, 1994). Over the past decade, a growing subset of studies has focused on the decomposition of active manager performance (Kacperczyk *et al.*, 2014;

Berk and Van Binsbergen, 2015), with many acknowledging the presence of a minority of skilled active managers that generate consistent performance alpha. At the same time, there has been a rising interest in incorporating ML and predictive modelling in the study of financial time-series forecasting (Krollner *et al.*, 2010), portfolio optimisation (Ban *et al.*, 2018), as well as stock selection strategies (Fu *et al.*, 2018; Rasekhschaffe and Jones, 2019).

Ludwig and Piovoso's (2005) seminal study applied supervised ML algorithms to predict the performance of US equity managers. The study utilised decision trees, feed-forward neural networks and naïve Bayes algorithms to map categorical features, namely relative quartile-based performance, tracking error and manager style to excess return labels measured over a two-year period. It is important to note that the study only covers the period January 1996 to December 2002, where data for the first five years (1996–2000) is reserved for training and the remaining two years (2001–2002) is used for testing. Results indicated that ML models trained on performance attributes are superior to a simple alpha momentum classifier. However, the inclusion of style categories negatively impacted model accuracy, barring that of the naïve Bayes algorithm. Regardless of the attribute set applied, the study confirmed that ML outperforms naïve selection methods, as ML models maintain aggregate precision scores exceeding 65%, compared to 61% achieved via the momentum classifier.

Wu *et al.*'s (2021) study applies ML models to sort hedge fund managers into portfolios based on predicted future performance. The authors use supervised ML models to estimate predictions using fund-specific performance attributes and economic variables as features, which are mapped to future monthly cross-sectional excess returns. The authors then sort the hedge funds into one of ten equally weighted decile portfolios based on predicted excess return and calculate post-ranking returns on a buy-and-hold basis. The study shows that the highest predicted return portfolio (decile 10) performs best out-of-sample, irrespective of the ML model applied. However, top-decile portfolios tend to exhibit significantly higher levels of volatility. The study also finds that return-based features produce the highest levels of out-of-sample nominal return. Lastly, across the ML models applied, those that are more adept at modelling non-linearities, such as random forest (RF), gradient boosting and deep neural nets, performed best.

DeMiguel *et al.* (2021) conducted a study applying several ML models to a cross-section of US actively managed mutual funds, mapping prior year fund characteristics to future annual risk-adjusted alpha. The authors ranked mutual funds into equally weighted deciles based on the predicted performance, and out-of-sample performance was measured over the following calendar year. Pooled OLS regressions were also used to rank managers into portfolios for comparable naïve returns, in order to test the potential non-linear benefits of the machine-learning models applied. Similar to Wu *et al.* (2021), the study found that ML models that are able to capture non-linear relationships in and between features and labels, such as RF and gradient boosters, are superior compared to parametric (linear) manager selection methods. The authors emphasise the necessary dynamism required in applying ML models, given that feature importance varies dramatically over the sample period. They believe that this finding fits neatly within an ML paradigm given the inherent flexible mapping afforded by ML algorithms and the added benefit of capturing non-linear relationships between features and labels.

Li and Rossi (2020) applied ML models to select mutual funds, focusing on fund holdings and style exposures as features. The authors suggest that the relationship between fund performance and factor exposure is non-linear, highlighting the advantage of ML over parametric methods for predicting fund performance. The study employs 94 share characteristics related to fund holdings, using ML models to predict excess returns over a single month. Mutual funds were then sorted into decile portfolios based on their predicted risk-adjusted return. Like previous studies, including Wu *et al.* (2021) and DeMiguel *et al.* (2021), funds with the highest predicted performance outperform their low predicted performance counterparts by a significant margin. The authors found that boosted regression trees (BRT) performed best in identifying future exploitable performance,

attributing this to BRT's ability to capture non-linearities and interactions between factor loadings and manager performance. These results confirm the benefit of ML methods over parametric linear frameworks in predicting mutual fund performance.

In contrast to the findings of Li and Rossi (2020), Kaniel *et al.* (2022) contend that fund characteristics and market state are more significant in predicting active fund risk-adjusted performance than share level characteristics. By employing feed-forward neural networks as their primary ML technique, Kaniel *et al.* (2022) identified that fund-level momentum, fund flow and market state are the most critical predictors. Additionally, when managers are classified into decile portfolios based on their ML predicted excess returns, the results demonstrate considerably positive long-only and long-short excess returns when compared to the lowest predicted return decile. Remarkably, out-of-sample performance is maintained even when share-specific characteristics are removed, contradicting the findings of Li and Rossi (2020). The study, however, validates the existing literature indicating that the benefits of ML mainly arise from the ability to detect and exploit non-linear interactions between predictors and excess returns. Table 1 below outlines the pertinent research that applies ML to anticipate active manager performance, taking into account the manager dataset, ML algorithms, feature and label variables used and critical findings.

3. Data and methodology

Our empirical framework involves several steps. Firstly, we sort managers into decile portfolios based on ML estimated predicted performance and calculate nominal portfolio returns. We then use naive manager performance predictors to conduct a comparative analysis of the different deciles. To ensure the robustness of our results, we repeat the analysis using risk-adjusted returns. We then evaluate the performance of ML algorithms by examining their out-of-sample predictive accuracy. This allows us to determine the effectiveness of ML algorithms compared to traditional scoring models in predicting manager performance.

3.1 Data

The fund manager data used in this study were obtained from Morningstar and consist of a cross-section of long-only South African equity managers that meet the criteria set by the Association for Savings and Investments South Africa (ASISA) South Africa General Equity classification. Both retail and institutional collective investment schemes (CIS) were included in the dataset. The data were collected on a monthly basis from January 2002 to December 2021 and included information on total returns net and gross of expenses, expense ratios, fund size and age.

To ensure the accuracy and reliability of the data, several adjustments and screens were applied. To mitigate the impact of survivorship bias, funds that exited the database were still retained but were no longer included in subsequent portfolio sorts. Funds were also assigned a single first true name to minimise the effect of structural breaks related to name changes. For funds with several fee classes, the fund fee class with the longest return series was retained. If fee class time-series were of equal length, preference was given to the lowest fee class. Additionally, funds with less than 36 months of continuous return data were excluded to avoid incubation bias. Finally, passive funds and Exchange Traded Funds (ETFs) were excluded, as the focus of this study is active manager performance.

After adjustments, the cross-section of managers decreased from 388 to 300, which is significantly smaller than the global samples used in international literature. However, there is still adequate scope for the sorting of portfolios, as the average number of qualifying managers over the period was 180, varying between a low of 120 and a high of 221.

Market capitalisation and factor-weighted style data were obtained from 27four Investment Managers and are used in addition to fund manager data. The style data,

Authors	Data	Features	ML	Finding
Ludwig and Piovoso (2005)	Cross-section of large and small capitalisation US equity managers over the period January 1996–December 2002	9 fund performance characteristics	Naïve Bayes, feed-forward neural net and decision trees	ML algorithms significantly outperform naive scoring model for large and small capitalisation US equity managers in terms of precision
DeMiguel *et al.* (2021)	Cross-section of US mutual funds over the period January 1980–December 2018	8 fund level characteristics and 7 fund performance measures	Elastic net, gradient boosters, random forests and feed-forward neural nets	All ML algorithms produce positive alphas while flexibility in feature importance is key driver of alpha generation
Wu *et al.* (2021)	Cross-section of US hedge funds over period January 1994–December 2015	17 fund performance characteristics and 4 macro-economic features	Lasso, random forest, gradient boosting and feed-forward neural net	ML-based sorts are more effective in forecasting future hedge fund returns than basic linear models with neural nets being the most effective. Additionally, return (performance-based features are superior in predicting future performance. Lastly, ML sorts are superior in terms of nominal returns at the expense of increased volatility
Li and Rossi (2020)	Cross-section if US equity mutual funds over the period January 1980–December 2018	94 fund holding level characteristics based on share characteristics resulting in factor exposure on a fund level	Lasso, elastic net, random forest, gradient boosting and feed-forward neural nets	Sorting funds on gradient boosting predictors results in significant alpha ad superior to pure linear models as well as univariate sorts. Of the top ten features, risk-based measures dominate the top 7 while momentum-based measures the remaining 3
Kaniel *et al.* (2022)	Cross-section of US mutual funds over period January 1980–January 2019	46 stock characteristics, 13 fund characteristics, investor sentiment and fund flows	Feed-forward neural nets	Fund characteristic information is main driver of future return, specifically fund flow and fund momentum
Chang *et al.* (2022)	Cross-section of US actively managed equity funds over the period November 2011–December 2018	Morningstar backward-looking star rating, analyst rating and quantitative rating based on random forest	Random forest	Morningstar analyst ratings are positively related to future alpha while random forest based quantitative rating is a poor predictor of mutual fund performance

Source(s): Table by authors

Table 1. Summary of pertinent literature focusing on ML and active manager performance prediction

which is detailed in Table A1, covers both long-only and style premium monthly returns and includes momentum, value, growth, low volatility, quality, size, currency risk, dividend yield and liquidity estimated using the top 100 shares by market capitalisation on the JSE. The style data spans from January 2000 and is updated monthly and is used to estimate several of the performance-based features described in the following section.

3.2 Research design – features and labels

The research design employed in this study is largely consistent with the international literature, taking inspiration mainly from Wu *et al.* (2021) and DeMiguel *et al.* (2021). We evaluate a total of 36 performance-related features that are measured over consecutive 60-month rolling windows, which broadly capture various aspects of fund performance, including historical cumulative performance, performance consistency, stylised alpha, information ratio, tracking error, as well as the proportion of illiquid and small-cap holdings. Table 2 that follows provides a detailed list of features applied.

As in DeMiguel *et al.* (2021), we apply annual as opposed to monthly future performance as the dependent/label variable where performance is proxied by excess return over the benchmark proxy. In terms of the calculation of excess return, a unique consideration specific to the study of South African active managers is the lack of a singular benchmark applicable across managers in the sample. At the time of writing, the majority of managers (82%) are benchmarked against the JSE Capped SWIX index (J433), a shareholder weighted index that limits the proportion in Naspers (JSE:NPN) to 10%. The benchmark was established in 2016 as Naspers maintained an aggregate weight of more than 25% in the JSE All-share index (J203) and JSE SWIX index (J403) [1]. Given the lack of consistent benchmarking

Feature group	Feature description	Variable type
Cumulative historical return	Cumulative returns over the most recent 1–5 years	Continuous
Style alpha	intercept of time-series regression of manager returns on style factor premiums	Continuous
Information ratio	Excess return over tracking error using - Full and Capped SWIX (J403/J433)[1] - Peer average - Style benchmark - Style grouped peer benchmark	Continuous
Performance consistency	Proportion of returns over 60-month window above benchmark, where benchmark proxied by - Full and Capped SWIX (J403/J433) - Peer average - Style benchmark - Style grouped peer benchmark	Continuous
Time-weighted performance consistency	Proportion of returns above benchmark where time-period is linear weighted to reduce impact of early observations and upweight more recent observations. Benchmarks applied identical to conventional consistency	Continuous
Best style	Is manager style best performing style over 60-month period	Binary
Small/illiquid holding	Combination of time-series regression coefficients run using SMB (small minus big) and IML (illiquid minus liquid). If combined coefficient is in top quartile, manager is assumed to have higher illiquid and small capitalisation holding	Binary

Table 2.
Feature descriptions

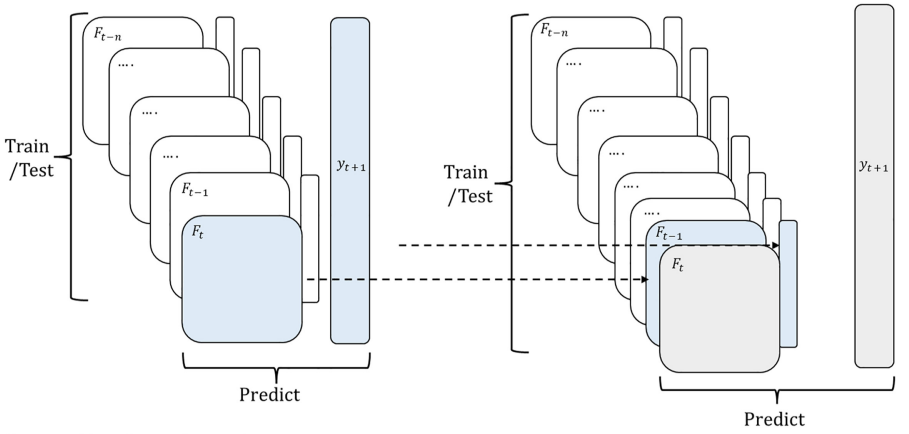
Source(s): Table by authors

cross-sectionally and through time, we opt for applying the ASISA SA General Equity category average as the primary benchmark proxy for calculating future 12-month excess performance.

Consistent with Wu *et al.* (2021), we adopt a cross-sectional framework instead of forecasting future time-series returns. However, we deviate from the literature by transforming the ML task from a regression problem to a classification problem [2]. The main aim of the literature is to rank managers based on predicted returns using ML. By employing a classification framework, ML models can forecast probability scores, allowing for manager ranking based on probability without the need to forecast future time-series excess returns. Therefore, we convert future excess returns over the subsequent year into a binary variable equal to one when cumulative excess returns are positive and zero otherwise, instead of using time-series or cumulative excess returns as labels. There are several benefits to this methodological deviation. Firstly, classification of a binary outcome is less complex for ML models on a like-for-like basis, particularly when compared to time-series forecasting. Secondly, the aim of the ML model is identical to the research objective of the paper, which is to identify managers who will either outperform or underperform. On the other hand, forecasting continuous variables like excess returns to sort managers is an unnecessary abstraction from the research objective since classification can achieve the same outcome.

3.3 Portfolio sorting procedure

The portfolio sorting method used in this study follows the approach of DeMiguel *et al.* (2021) and Wu *et al.* (2021), where only historical features and forward excess performance are employed for model training, testing and prediction. The features and labels are measured over rolling historical 60-month and forward 12-month windows, resulting in a total of 15 feature matrices and label vectors for the entire sample period. The process is illustrated in a diagram provided in Figure 1. To ensure an adequate cross-section of features, the initial training involves using nine years' worth of features from 2002 to 2010 (white blocks). At each portfolio sort date, the model is trained on historical feature data excluding the most recent 60-month feature set. The model is then applied to the most recent feature set (unseen by the model) to predict manager performance over the following 12 months.



Source(s): Figure by authors

Figure 1.
Train, test and predict procedure applied

For instance, the 2011 predictions (blue block Yt+1) and manager rankings are based on the parameters trained on features and labels spanning 2002 to 2009 (white blocks) and predicted using the most recent features measured over 2006 to 2010 (blue feature matrix F_t). Managers are then sorted into equally weighted decile portfolios based on their probability rank, and their performance is tracked on a buy-and-hold basis for the next year, after which the process is repeated. After each portfolio sort date, the most recent feature and label data (related to 2011) is added to the training data, and an updated model is used to forecast probabilities for 2012 (grey block Y_{t+1}) using the feature set measured over 2007 to 2011 (grey block F_t).

3.4 ML models

In line with prior literature, we employ six ML models to predict active manager performance: lasso, elastic-net, support vector machine, random forest, XG and LG. Our focus is not on identifying the optimal model, but rather on examining whether ML techniques can be used to predict returns of active managers. Thus, our aim is to use these models to classify managers based on their future performance by estimating the probability of under or outperformance using a general forecast classification model as described in equation 1 below

$$\widetilde{P}^{i}_{t+q} = \frac{1}{1e^{-g_t(\Omega^{i}_{t})}} \tag{1}$$

Where $\widetilde{P}^{i}_{t+q}$ is the probability of excess performance for fund i over q future periods being positive, $g_t(.)$ is a forecast function that can be either linear or non-linear and Ω^{i}_{t} is a matrix of features or fund i characteristics known at time t.

We employ a total of six models in our study to predict active manager performance. These models comprise of three linear and three non-linear models and are selected based on prior literature. The linear models used are Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996), elastic net (EN) by Zou and Hastie (2005) and support vector machines (SVM) introduced by Cortes and Vapnik (1995). LASSO is a regression method that uses L_1 regularisation by adding a penalty equal to the absolute value of the fitted coefficients, resulting in a more parsimonious model through parameter elimination. Elastic net, on the other hand, combines L_1 and L_2 penalties to add a quadratic component to the penalty function, making it a better alternative to LASSO. SVM, a highly popular classification algorithm, relies on a hyperplane (n-1 dimensional subset of a n-dimensional Euclidian space) to divide data into classes by maximising distance (margin) between support vectors to define the optimal hyperplane. SVM is known for its efficiency and speed, and is often used in a wide range of classification tasks.

In our study, we also include three non-linear models, namely random forest (RF), XG and LG, to capture and model non-linear interactions between features and labels. RFs, per Breiman (2011), are an ensemble method that applies decision trees and corrects for overfitting by using bootstrap aggregation or "bagging" to the tree-learners. Bagging involves the repeated random sampling of the training set and fitting trees, which improves model performance by reducing variance and enhancing out-of-sample generalisation.

XG and LG both apply forms of gradient boosting per Friedman (2002), and they differ in terms of their tree growth assumptions. The former allows trees to grow depth-wise while the latter grows trees leaf-wise. Both models apply a decision tree framework, but they differ from RF as they rely on the addition of multiple weak learner decision trees in a stage-wise approach on an iterative basis. One disadvantage is that the iterative process can be applied ad infinitum in-sample, leading to overfitting and poor out-of-sample generalisation, but this can be controlled through regularisation techniques such as minimising the number of iterations applied, tree depth and increasing learning rates.

3.5 Hyper-parameter tuning

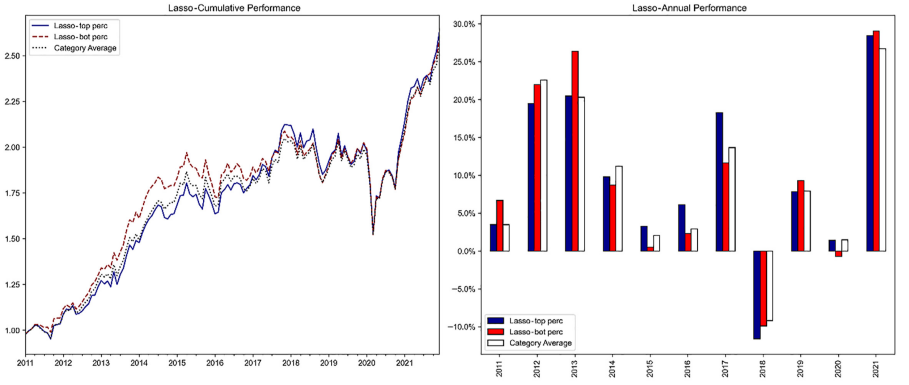
In our study, hyper-parameter tuning is employed to determine the best set of hyper-parameters for each model, which allows us to achieve the highest level of accuracy while minimising the risk of overfitting. Following the methodology of DeMiguel *et al.* (2021), we use k-fold cross-validation on the training data, assuming $k = 5$, and at each portfolio sort period, we initially divide the training data into a 70/30 train-test split. Then, we partition the training data into k folds, reserving one group for validation purposes and using the others to train the model. The hyper-parameters for each model are then applied across the *k-1* instances, resulting in the training and evaluation of 12 models ((5–1) *x* 3) against 3 label validation groups, assuming three hyperparameters are considered. The cross-validation score is saved in a "grid," and the optimal model is selected based on the set of hyper-parameters that achieved the highest cross-validation accuracy score. The optimised model is then applied to the test set (30%), and its accuracy score is compared against an untuned base model. If the precision is improved, the "tuned" model is implemented.

As shown in Figure 1, the training set is expanded at each portfolio sort date to incorporate the previous feature matrix and label vector. Consequently, hyper-parameter tuning is performed at each portfolio sort date, meaning that each sort relies on a new model that is trained (and optimised) on the most up-to-date available data and applied to unseen data. Finally, following the approach of DeMiguel *et al.* (2021), we allow for randomisation of the training data, as Coulombe *et al.* (2022) demonstrate that k-fold cross-validation is a best practice when applying ML to panel (time-series tabular) data (Refer to Table A2, which provides information on the various hyper-parameter grids applied to each ML model, Table A3 for the combined train and test precision scores, Table A4 for the out-of-sample precision scores, and Table A5 for the total feature importance for each model over the study period).

4. Results

4.1 Decile performance

Figures 2–7 present a graphical analysis of the out-of-sample performance of the top and bottom decile manager portfolios across the 6 ML models employed in this study. The results reveal that the performance of the linear models, including LS and EN, is largely unimpressive, as they produce top and bottom decile portfolios that track the category average. Meanwhile, the extreme deciles of the SVM model underperform the category average.



Source(s): Figure by authors

Figure 2.
Lasso cumulative and discrete time-period performance

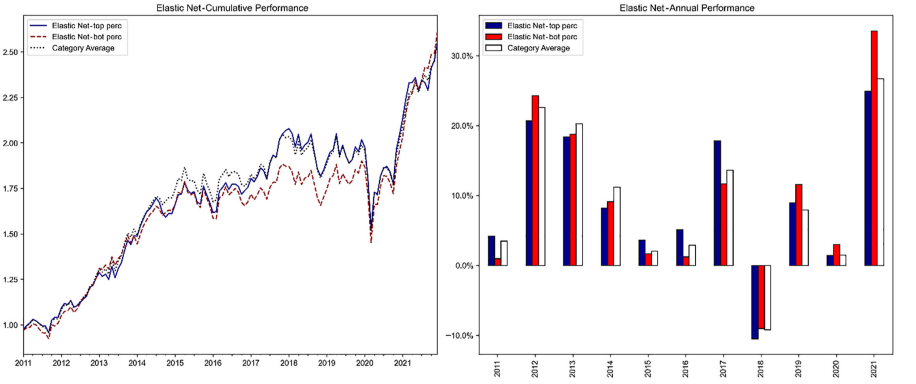


Figure 3.
Elastic net cumulative and discrete time-period performance

Source(s): Figure by authors

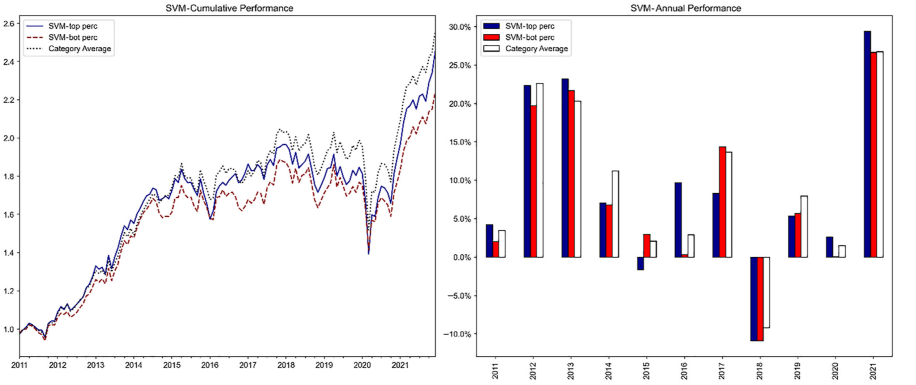


Figure 4.
SVM cumulative and discrete time-period performance

Source(s): Figure by authors

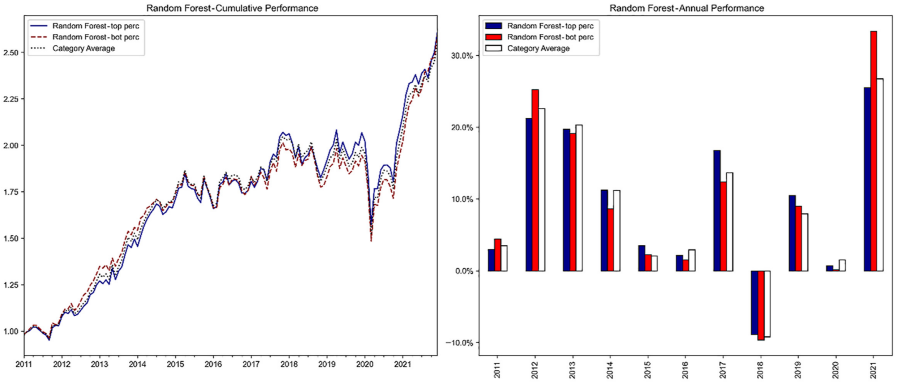
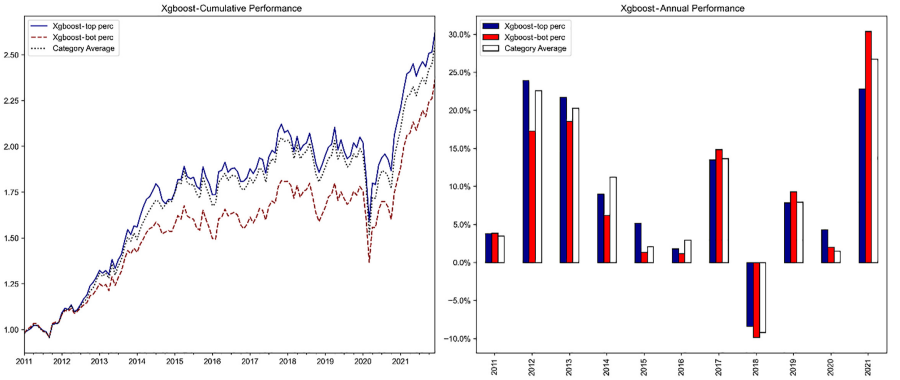


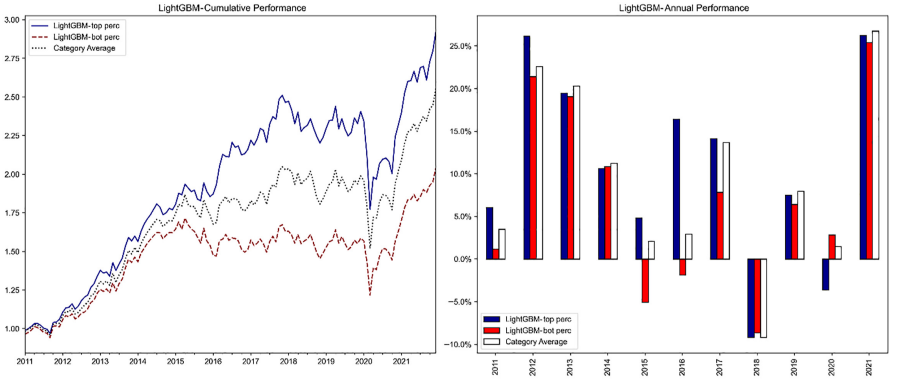
Figure 5.
Random forest cumulative and discrete time-period performance

Source(s): Figure by authors



Source(s): Figure by authors

Figure 6. XGBoost cumulative and discrete time-period performance



Source(s): Figure by authors

Figure 7. LightGBM cumulative and discrete time-period performance

A closer examination of the discrete period performance reveals that the LS and EN top decile portfolios produce almost identical annual returns from 2012 to 2020. On the other hand, the SVM top decile outperforms the bottom decile over the period of 2011–2014, 2016 and 2020. However, in comparison to both the bottom decile and the category average, the SVM top decile performs the worst in 2015, 2017 and 2019.

Figures 5–7 present the extreme decile portfolios sorted using ML algorithms that are more effective in capturing non-linear relationships. RF produces moderate outperformance in the top decile portfolio and underperformance for the bottom decile portfolio. The former outperforms the latter in eight out of the eleven-year evaluation period (2013–2020), but only outperforms the category average in five years (2014–2015, 2017–2019). In contrast, gradient boosting algorithms demonstrate a significant improvement in both relative and absolute performance, as shown in Figures 6 and 7. The top and bottom decile portfolios sorted via XG exhibit a substantial performance spread, with marginal cumulative outperformance for the former and considerable underperformance for the latter. The discrete performance figure reveals that the top decile outperforms the bottom decile in seven years of the eleven-year period, while the bottom decile underperforms the category average in six years. It is noteworthy that much of the underperformance of the bottom decile portfolio occurs over the 2012–2016 period. Conversely, the latter half of the evaluation period shows a significant

improvement in the bottom decile's performance, achieving returns in excess of both the top decile and category average in 2017, 2019 and 2021.

The performance of extreme decile portfolios sorted using LG provides the most consistent results with the current body of developed market literature. The top decile portfolio exhibits cumulative out-of-sample performance well above the category average, while the bottom decile shows the lowest cumulative performance across all ML algorithms considered. This result is also confirmed through the discrete period performance analysis, with the top decile significantly outperforming the bottom decile in eight out of eleven years, particularly in 2012 and 2015 to 2017.

Tables 3–5 report the return statistics for the top and bottom decile manager portfolios across the ML models used, as well as the category average as the market proxy for

Table 3.
Table describes performance metrics for the predicted top-decile manager portfolios across the machine learning models applied

Model	Annual return	Sharpe ratio	Information ratio
Lasso	9.58%**	0.087	0.050
Elastic net	9.29%**	0.081	0.003
SVM	8.94%**	0.071	−0.039
Random forest	9.48%***	0.086	0.035
LightGBM	10.52%***	0.111*	0.111
XGBoost	9.51%***	0.088	0.037
Category average	9.26%**	0.082	N/A

Note(s): Returns are evaluated statistically using paired sample t-tests against the risk-free rate proxied by the South African government 91-day t-bill while Sharpe ratio p-values are estimated applying the methodology of Bailey and Lopez de Prado (2012). Sharpe ratios are calculated applying the top-decile portfolio return in excess of the risk-free rate scaled by standard deviation while information ratio applies excess return over the category average scaled tracking error. *,**,*** indicates significance at the 10%, 5 and 1% level
Source(s): Table by authors

Table 4.
Table describes performance metrics for the predicted bottom-decile manager portfolios across the machine learning models applied

Model	Annual return	Sharpe ratio	Information ratio
Lasso	9.39%***	0.085	0.020
Elastic net	9.48%***	0.086	0.029
SVM	8.02%**	0.053	−0.163
Random forest	9.43%***	0.085	0.021
LightGBM	7.23%**	0.032	−0.273
XGBoost	8.57%**	0.065	−0.127
Category average	9.26%**	0.082	N/A

Note(s): Returns are evaluated statistically using paired sample t-tests against the risk-free rate proxied by the South African government 91-day t-bill while Sharpe ratio p-values are estimated applying the methodology of Bailey and Lopez de Prado (2012). Sharpe ratios are calculated applying the bottom-decile portfolio return in excess of the risk-free rate scaled by standard deviation while information ratio applies excess return over the category average scaled tracking error. *,**,*** indicates significance at the 10%, 5 and 1% level
Source(s): Table by authors

Table 5.
Table describes the zero-cost long-short predicted top decile minus bottom decile manager portfolios over the period January 2011 to December 2021

Top-bottom	Lasso	Elastic net	SVM	Random forest	LightGBM	XGBoost
Excess return	0.19%	−0.20%	0.93%	0.05%	3.29%	0.94%
P-value	0.8184	0.8433	0.3609	0.9554	0.0107**	0.1773

Note(s): Excess returns are expressed annually, and *p*-values are calculated applying a paired sample *t*-test
Source(s): Table by authors

comparison. In Table 3, LG produces the highest annualised return of 10.5% per annum for the long-only top-decile portfolio, which is statistically significant at the 1% level compared to the risk-free proxy. In contrast, the SVM algorithm produces the worst-performing top-decile portfolio, with an annualised return of 8.9%, which is 0.3% lower than the category average. When considering both return and portfolio risk, the LG top-decile portfolio provides the highest monthly Sharpe and Information ratios [3].This equates to 11 basis points of excess return per unit of absolute volatility and active risk (tracking error), respectively. In contrast, the SVM top-decile portfolio produces the lowest Sharpe and Information ratios.

In Table 4, the performance results for the bottom-decile portfolios are presented. Notably, three out of 6 ML models, namely LS, EN and RF, produce bottom-decile portfolios that outperform the category average in terms of annual return and monthly Sharpe ratio. However, none of the Sharpe ratios are statistically different from zero. The worst-performing bottom-decile portfolio is produced by LG, which underperforms the category average by an average of −2% per annum. The SVM and XG bottom-decile portfolios are the second and third worst performers, respectively, underperforming the category average by −1.3% and −0.7% per annum, respectively.

Table 5 presents the excess returns of the long-short top-minus-bottom decile portfolios. The results show that only LG produces a significant excess annual return of 3.3%, indicating that it can effectively distinguish between the best and worst performing managers in the market. XG achieves the second-best annualised excess return of 0.9% but misses significance at the 10% level, suggesting that its performance in identifying winning and losing managers is less precise than LG. The LS algorithm performs well in selecting top-decile managers, outperforming the category average by 0.32% per annum, but achieves an insignificant long-short excess return of 0.2% over the bottom decile, suggesting that it lacks precision in predicting manager underperformance. Overall, the results highlight the importance of precision in distinguishing future winning and losing active managers when constructing long-short portfolios.

4.2 Naïve manager selection

As a point of comparison, this study applies naïve manager selection methods to the data over the sample period, based on established literature. Momentum, as defined by Jegadeesh and Titman (1993), is proxied by 12-month cumulative return. Long-term reversal, as defined by De Bondt and Thaler (1985), is proxied by 60-month cumulative return. Lastly, style alpha is estimated via time-series regression, whereby excess manager returns are regressed on long-short style premiums over a historical 60-month window. These methods are applied to assess the relative performance of ML models in identifying skilled emerging market active equity managers.

Figures 8–10 depict the outcomes of the portfolio sorts based on the naïve selection methods against the category average, both on a cumulative and discrete period basis. As shown in Figure 8, the momentum top-decile portfolio exhibits superior performance compared to both the category average and bottom-decile portfolio on a cumulative basis. Additionally, in terms of discrete annual performance, it outperforms the bottom-decile and category average in nine- and seven years, respectively, during the eleven-year assessment period. On the other hand, the long-term reversal top-decile (long-term loser managers) portfolio tracks the category average, while the long-term winners (bottom-decile) portfolio outperforms both the category average and top-decile portfolios. Figure 10 presents that the style alpha top-decile portfolio surpasses the category average and bottom-decile portfolio. However, this outcome is primarily driven by positive excess performance since 2019, with the previous performance providing minimal alpha generation or divergence from the category average.

Tables 6 and 7 present performance statistics on the top and bottom-decile manager portfolios sorted on momentum, reversal and style alpha. In Table 6, both the momentum and style alpha top-deciles outperform the category average, achieving annualised returns of 10.5

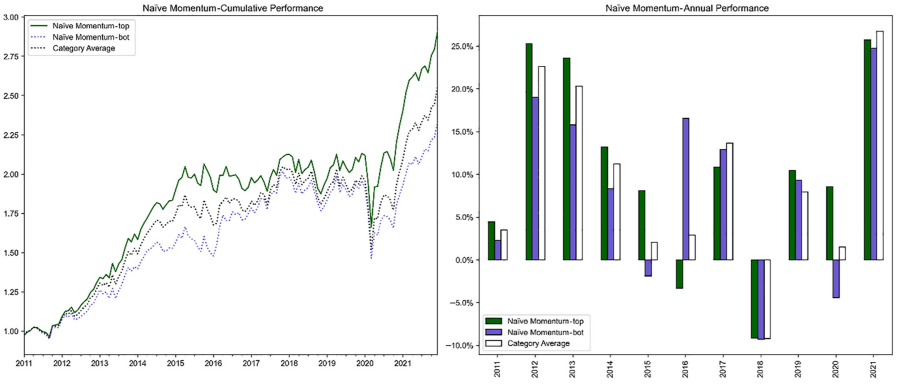


Figure 8.
Naïve manager momentum cumulative and discrete period performance

Source(s): Figure by authors

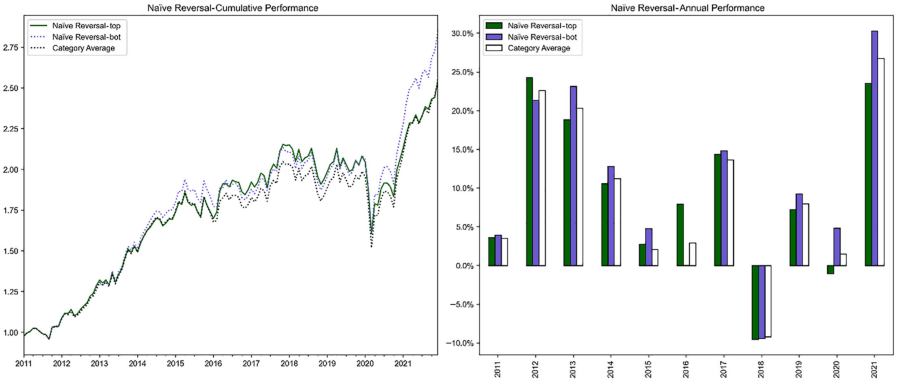


Figure 9.
Naïve manager reversal cumulative and discrete period performance

Source(s): Figure by authors

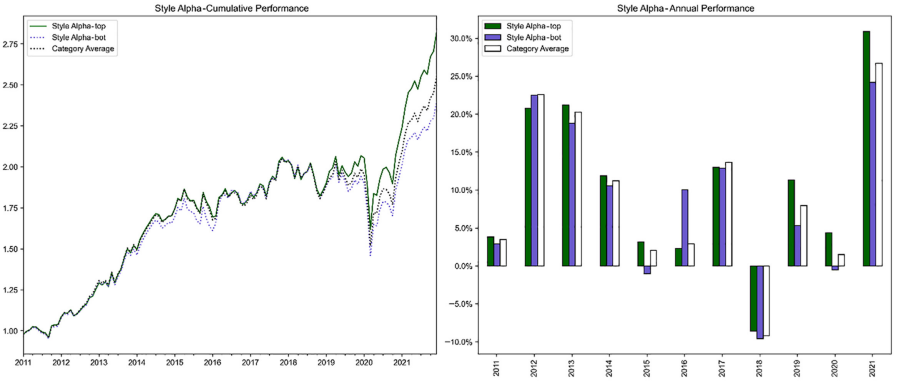


Figure 10.
Style alpha cumulative and discrete period performance

Source(s): Figure by authors

and 10.2% per annum on average respectively. Momentum produces a slightly higher (but statistically insignificant) Sharpe ratio, while style alpha produces the highest information ratio. In comparison to the ML top-decile portfolios, naïve methods perform better, with the exception of LG. On the other hand, Table 7 shows that, except for long-term reversal, both momentum and style alpha produce bottom-decile portfolios that underperform the category average.

Table 8 presents the annualised top minus bottom decile excess returns of the momentum, reversal and style alpha sorted portfolios. The momentum portfolio generates the highest excess return of 2% per annum, but the result is statistically insignificant. Conversely, the style alpha portfolio produces a lower but significant (at the 5% level) excess return of 1.5% per annum. These findings suggest that performance momentum and style alpha may be useful in predicting future performance of South African active equity managers. Additionally,

Top decile	Annual return	Sharpe ratio	Info. ratio
Naïve momentum	10.45%***	0.110	0.157
Naïve reversal	9.16%***	0.083	−0.020
Style alpha	10.16%***	0.105	0.221
Category average	9.26%**	0.082	

Note(s): Returns are evaluated statistically using paired sample t-tests against the risk-free rate proxied by the South African government 91-day t-bill while Sharpe ratio p-values are estimated applying the methodology of Bailey and Lopez de Prado (2012). Sharpe ratios are calculated applying the naïve top-decile portfolio return in-excess of the risk-free rate scaled by standard deviation while information ratio applies excess return over the category average scaled tracking error. *,**,*** indicates significance at the 10%, 5 and 1% level
Source(s): Table by authors

Table 6. Table describes annualised performance metrics for the predicted top-decile manager portfolios applying naïve selection criteria, namely momentum, reversal and style alpha

Bottom decile	Annual return	Sharpe ratio	Info. Ratio
Naïve momentum	8.44%**	0.062	−0.083
Naïve reversal	10.28%***	0.104	0.251
Style alpha	8.68%**	0.068	−0.106
Category average	9.26%**	0.082	

Note(s): Returns are evaluated statistically using paired sample t-tests against the risk-free rate proxied by the South African government 91-day t-bill while Sharpe ratio p-values are estimated applying the methodology of Bailey and Lopez de Prado (2012). Sharpe ratios are calculated applying the naïve bottom-decile portfolio return in-excess of the risk-free rate scaled by standard deviation while information ratio applies excess return over the category average scaled tracking error. *,**,*** indicates significance at the 10%, 5 and 1% level
Source(s): Table by authors

Table 7. Table describes annualised performance metrics for the predicted bottom-decile manager portfolios applying naïve selection criteria, namely momentum, reversal and style alpha

Top-bottom	Naïve momentum	Naïve reversal	Style alpha
Annual excess return	2.01%	−1.12%	1.47%
P-value	0.1563	0.1334	0.0381**

Note(s): Excess returns are expressed annually, and p-values are calculated applying a paired sample *t*-test
Source(s): Table by authors

Table 8. Table describes the zero-cost long-short predicted top decile minus bottom decile manager portfolios, applying naïve sorting criteria, over the period January 2011 to December 2021

the results indicate that the naïve manager selection methods (momentum and style alpha) perform better than ML methods on aggregate over the sample period. However, LG is still the superior ML algorithm in terms of long-only and excess return performance.

4.3 Risk-adjusted returns

Time-series attribution or factor spanning tests are a useful tool to determine risk-adjusted performance through alpha estimates and provide insights into the stylised factors driving portfolio performance. The following tables describe risk-adjusted performance for the ML-based top-decile portfolios using two models: a pure Carhart (1997) four-factor model and a modified Carhart (1997) model that replaces share-based momentum with manager momentum. Due to space constraints, we only present results for the top-decile portfolios.

Table 9 presents the results of the time-series attribution analysis using the Carhart (1997) four factor model. The LG top decile portfolio shows the highest alpha estimate of 0.14% per month, which is statistically significant at the 10% level. All portfolios show significant factor loadings on the market risk premium, which is proxied by the category average minus the risk-free proxy. This result is expected due to the limited investment universe of the JSE and the high level of benchmark cognisance among equity managers. The RF top decile portfolio leans towards managers with a large-cap bias, while the LG top decile portfolio tends to favour managers with a lower market capitalisation focus. The XG top decile portfolio exhibits a significantly negative factor loading on the value premium, indicating that the model prefers active managers with a growth bias. Additionally, only the LG top decile portfolio has a significant loading on the momentum premium. The negative sign of the loading implies that momentum or "hot-hands" does not explain LG's excess return.

To summarise, Table 10 provides the results of the same test as Table 9, but with manager momentum replacing price momentum. The LG top decile portfolio's alpha increases to 0.17% per month and becomes significant at the 5% level, and its momentum factor loading becomes nine times larger in absolute terms and is now significant at the 1% level. The market beta factor loadings are still large and significant, while the LG top decile shows a reduced bias towards small caps. The SVM top decile now shows a positive loading on the value premium, potentially explaining its underperformance over 2015–2019 (see Page *et al.*, 2022), while the XG top decile maintains a significantly negative loading, implying a growth bias. These results confirm that LG's superior nominal returns and significantly positive time-series alpha can be explained by a bias towards managers with negative recent momentum, who tend to experience a reversal in return.

	α	β_{MRP}	β_{SMB}	β_{VMG}	β_{WML}	Adj. R^2
Lasso	0.0003	1.0168***	−0.0089	0.0093	−0.0037	97.77%
Elastic net	0.0001	0.9939***	−0.0102	0.0159	−0.0091	96.74%
SVM	−0.0001	1.0086***	0.0115	0.0215	−0.0166	96.37%
Random forest	0.0002	1.0006***	−0.0192*	0.0080	0.0083	97.81%
XGBoost	0.0001	0.9967***	−0.0116	−0.0268**	−0.0025	97.62%
LightGBM	0.0014*	0.9588***	0.0431**	0.0198	−0.0284*	93.41%

Table 9.
Table describes time-series spanning regressions on the long-only top-decile manager portfolios generated using the six machine learning models applied

Note(s): The time-series regression takes the form of $R_{i,t} - r_f = \alpha_i + \sum \beta_{i,j} X_j + \varepsilon_t$ where manager returns in excess of the 91-day South African government t-bill rate are regressed against X_j, a matrix of time-series excess returns of the market proxy, size (small-minus-big), value (value-minus-growth) and momentum (winner-minus-loser). The spanning model therefore follows the specification of Carhart (1997), representing a Fama-French three factor model plus momentum. The table reports time-series regression coefficients that have been estimated applying Newey–West consistent standard errors with ***, **, * representing statistical significance at the 1%, 5 and 10% level

Source(s): Table by authors

4.4 Model accuracy

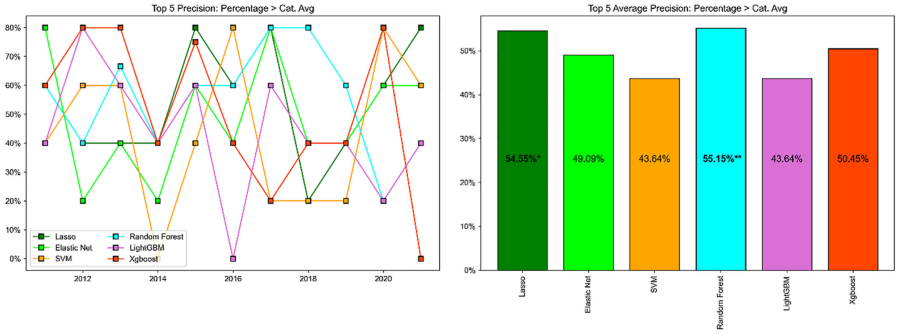
In order to evaluate the accuracy of the ML models used, two tests were conducted. The first focused on the predicted outperformers and tested the proportion of top managers per model that exceeded the category average over the following period, with the sub-sample of top managers increasing from five to twenty. The purpose of this test was to determine whether the top decile performance metrics presented were related to generalised model precision or rather specific manager performance. The second test evaluated the *ex ante* rank correlation of the best and worst predicted managers with the ex-post performance rank using spearman rank correlation. This test was beneficial in establishing whether the ML rank and subsequent performance rank maintained a monotonic relationship, thereby testing the ability, accuracy and consistency of each model in determining future top and bottom performing managers.

Figures 11–14 present the average accuracy of each ML model in predicting managers that outperform the category average. Accuracy is defined as the ratio of the number of predicted outperformers (k) that successfully outperformed the category average over the subsequent period (j), or $\frac{j}{k}$. In Figure 11, which considers the top five managers per model, LS and RF perform well, achieving precision rates of 54.6 and 55.2%, respectively. Both are statistically significant at the 10 and 5% levels. Moving to the top ten managers in Figure 12, RF's accuracy improves to 56.1% and becomes significant at the 1% level, while EN improves to 51.8%. However, LS's accuracy decreases to 50.9%. In Figures 13 and 14, which consider the top fifteen and twenty managers, LG achieves the highest accuracy of 55.2% (in Figure 13) and is

	α	β_{MRP}	β_{SMB}	β_{VMG}	β_{MWML}	Adj. R^2
Lasso	0.0003	1.0190***	−0.0113	0.0092	−0.0412	97.79%
Elastic net	0.0001	0.9954***	−0.0070	0.0199	0.0146	96.73%
SVM	−0.0002	1.0127***	0.0144	0.0273*	−0.0114	96.31%
Random forest	0.0002	0.9979***	−0.0190*	0.0059	0.0264	97.80%
XGBoost	0.0001	0.9975***	−0.0118	−0.0262**	−0.0086	97.62%
LightGBM	0.0017**	0.9744***	0.0271	0.0195	−0.2842***	94.34%

Note(s): The time-series regression takes the form of $R_{i,t} - r_f = \alpha_i + \sum \beta_{i,j} X_j + \varepsilon_t$ where manager returns in excess of the 91-day South African government t-bill rate are regressed against X_j, a matrix of time-series excess returns of the market proxy, size (small-minus-big), value (value-minus-growth) and **manager momentum** (winner-minus-loser managers). The spanning model differs to Table 4a as share level momentum has been replaced by manager level momentum, applying a 12-month look back and 1-month holding period. The table reports time-series regression coefficients that have been estimated applying Newey–West consistent standard errors with ***,**,* representing statistical significance at the 1%, 5 and 10% level
Source(s): Table by authors

Table 10. Time-series spanning regressions on the long-only top-decile manager portfolios generated using the six machine learning models applied



Source(s): Figure by authors

Figure 11. Model precision based on top 5 managers

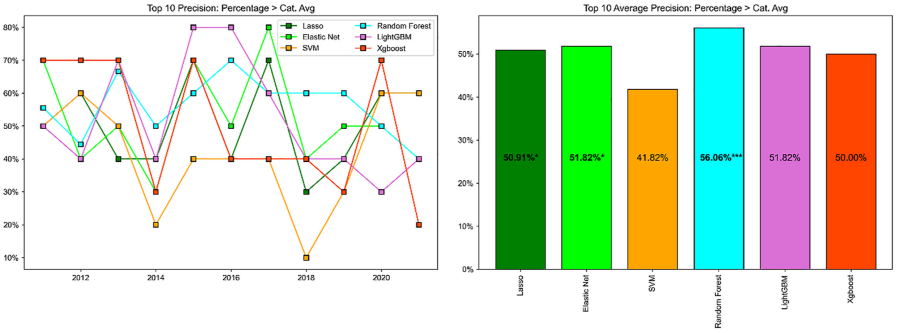


Figure 12.
Model precision based on top 10 managers

Source(s): Figure by authors

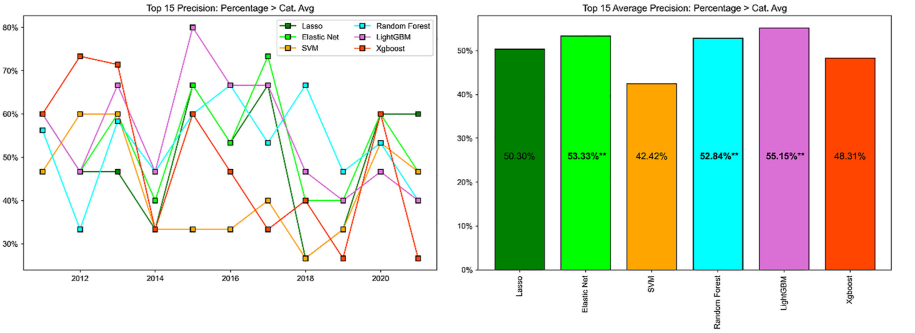


Figure 13.
Model precision based on top 15 managers

Source(s): Figure by authors

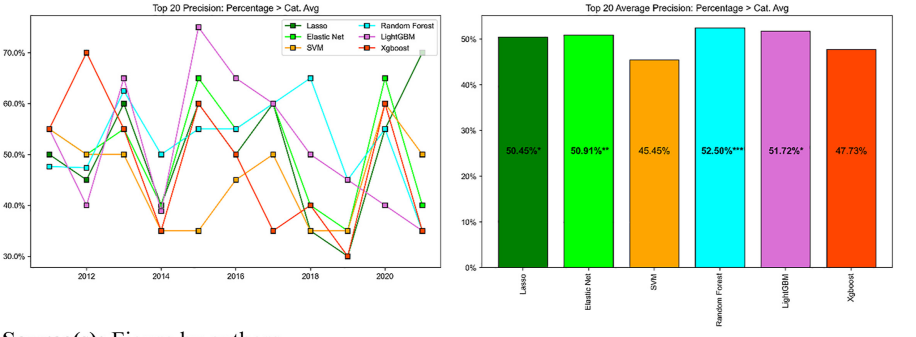


Figure 14.
Model precision based on top 20 managers

Source(s): Figure by authors

significant at the 5% level. RF maintains an average accuracy of over 52% in both, while EN also improves to 53.3% in Figure 13. Figure 14 shows that RF and LG are the best performers when considering the top twenty managers per model, achieving accuracy rates of 52.5 and 51.7%, respectively. RF is significant at the 1% level, while LG is significant at the 10% level.

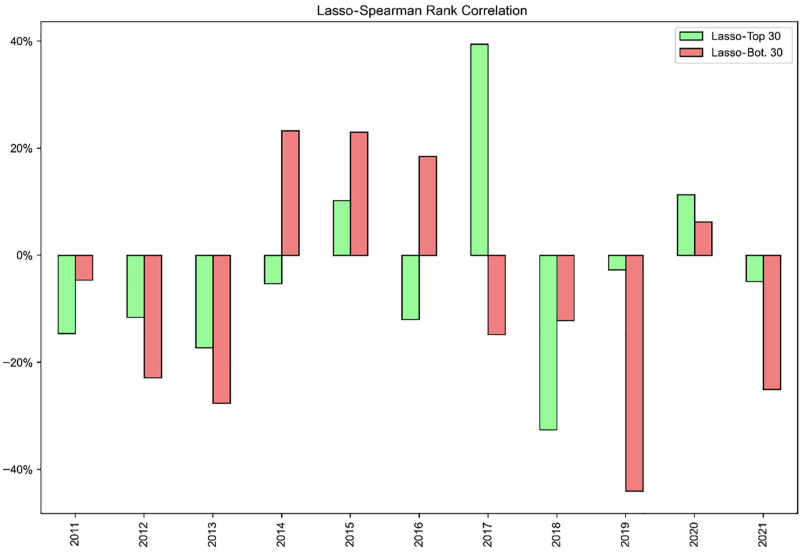
The accuracy tests reveal that RF, which performed poorly in the return-based tests, achieved the highest level of accuracy in terms of selecting managers that outperform the benchmark over the subsequent period. This finding is consistent with the literature on developed markets, as RF is part of the subset of ML models that can capture non-linear

relationships in data. LG, which did not perform well in the top five and ten manager tests, emerged as the second-best performing model in terms of accuracy, particularly when considering more than ten managers. However, it is important to note that these results should be viewed with caution, as the precision tests suggest that methodological design may drive performance outcomes. In this regard, the RF model, which performed well in terms of accuracy, did not deliver strong nominal and risk-adjusted returns, while LG, which was the best-performing model in investment performance, did not feature prominently in the accuracy tests for the top five and ten managers. One potential explanation for these results may be the application of decile portfolio sorts, which typically included between fifteen to twenty-five managers per portfolio during the sample period.

As a means of evaluating the accuracy and consistency of the ML models, we conduct a robustness test by assessing the rank correlation of the top and bottom predicted managers with their *ex-post* performance rank. This approach is aimed at determining the models' ability to identify future top and bottom performing managers. Specifically, we analyse the rank correlation of the top and bottom 30 predicted managers to establish whether the models are better at identifying these performers. The positive rank correlations would suggest that the models are more accurate in identifying future top and bottom performers.

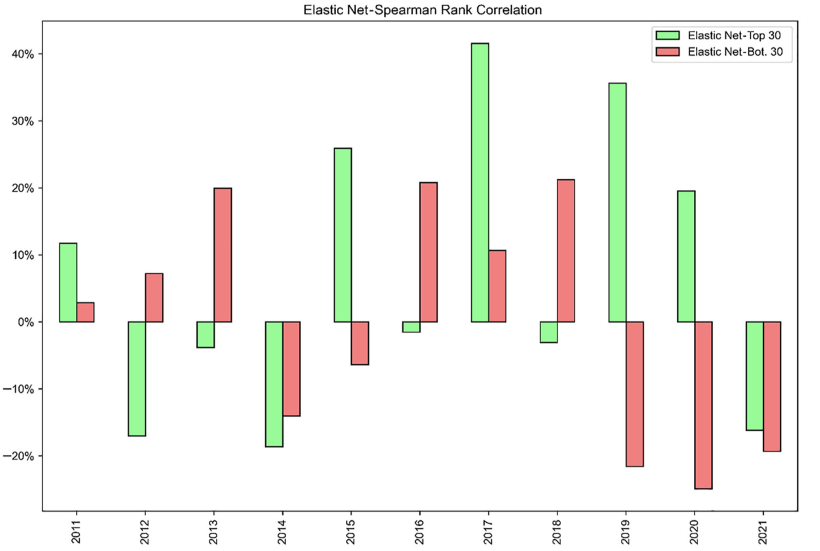
Furthermore, the rank correlation results can be interpreted alongside the investment performance results reported in Figures 2–7 to determine if there is a correlation between rank correlation and investment performance. This approach is expected to provide insight into whether the models' investment performance is consistent with their ability to identify future top and bottom performers.

In Figures 15–20, we present the results of the spearman rank correlation tests on the 6 ML models applied, where we compare the top (green bar) and bottom (red bar) 30 manager *ex ante* rank to their *ex-post* performance rank against the market proxy. Figure 15 shows the LS results and indicates that *ex ante* predicted rank maintains a positive relationship with ex-post manager performance rank for the top 30 managers in only two periods (2017 and 2020). In contrast, the LS model achieves positive rank correlations for the bottom 30 managers in four years, namely



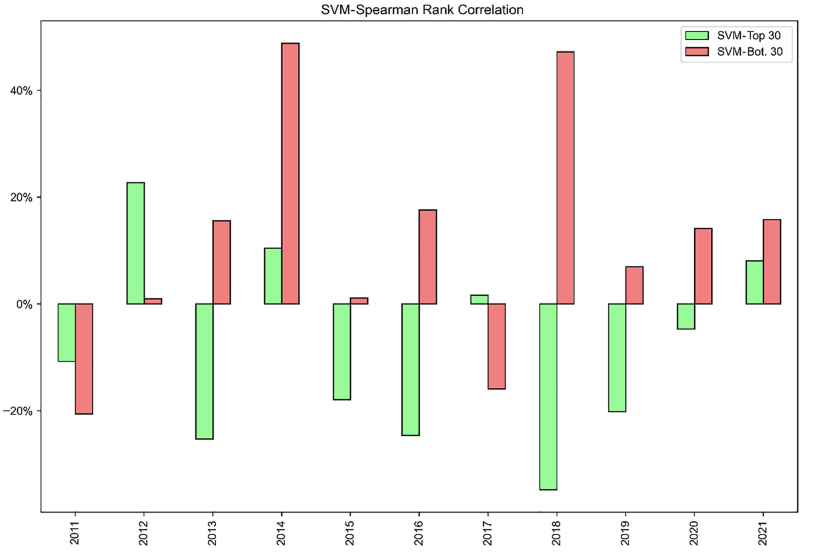
Source(s): Figure by authors

Figure 15.
Annual rank correlations between LS predicted top and bottom 30 manager performance against actual performance rank



Source(s): Figure by authors

Figure 16.
Annual rank correlations between EN predicted top and bottom 30 manager performance against actual performance rank

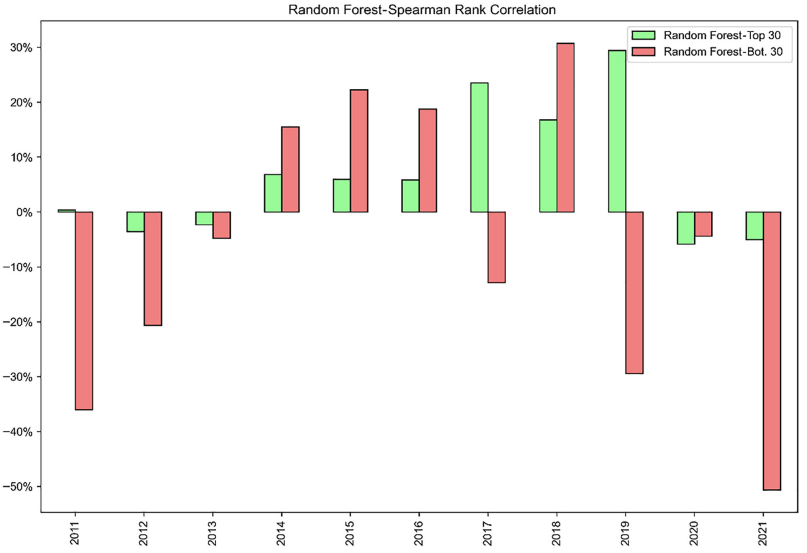


Source(s): Figure by authors

Figure 17.
Annual rank correlations between SVM predicted top and bottom 30 manager performance against actual performance rank

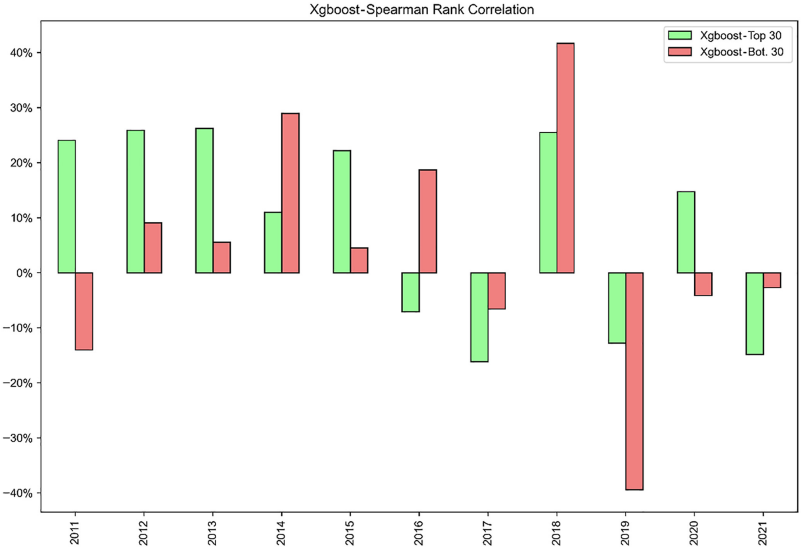
2014–2016 and 2020. When comparing Figures 15–20, it is observed that the top (bottom) decile achieves the highest (lowest) relative outperformance over the same years. Conversely, in periods where the correlation is negative, the LS model underperforms, namely in 2011–2013 and 2018.

The findings presented in Figure 16 indicate that the EN model exhibited positive rank correlations in five and six years for the top and bottom 30 managers, respectively. It is worth noting that the model was particularly effective at identifying top-performing managers



Source(s): Figure by authors

Figure 18. Annual rank correlations between RF predicted top and bottom 30 manager performance against actual performance rank



Source(s): Figure by authors

Figure 19. Annual rank correlations between XG predicted top and bottom 30 manager performance against actual performance rank

during 2019–2020, but less accurate in predicting manager underperformance. When comparing Figure 3 with Figure 16, it is evident that the top decile portfolio performed better than the category average in 2019 and 2020. However, the bottom decile achieved the highest annual returns over the same period, which is a direct reflection of the strong negative correlation between *ex ante* selection and *ex post* performance rank of the bottom 30 managers predicted by the EN model.

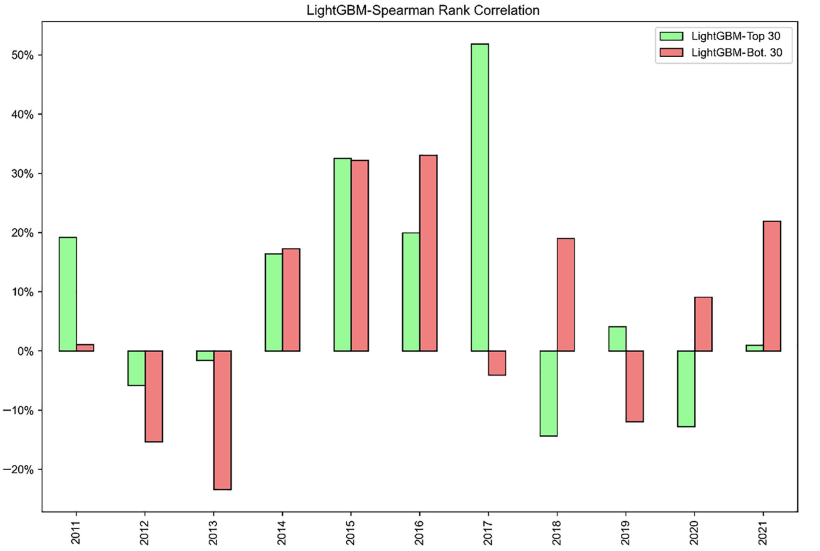


Figure 20.
Annual rank correlations between LG predicted top and bottom 30 manager performance against actual performance rank

Source(s): Figure by authors

In terms of nominal and risk-adjusted performance, the SVM model is the poorest performing ML algorithm on a long-only basis yet was successful in identifying future underperformance. As shown in Figure 17, the top 30 rank correlation is positive in only two years, 2012 and 2021. Conversely, for the bottom 30 managers, the *ex ante* rank positively correlates with the ex-post performance rank in seven years of the eleven-year period (nine years if counting 2012 and 2015 which are very close to zero). These results are consistent with the performance results presented in Figure 4, where the SVM model's bottom decile experiences substantial underperformance over the years 2013–2014, 2016 and 2018, all of which are periods of positive and relatively large rank correlations.

According to the findings presented in Figures 18–20, the RF, LG and XG models, which are known to be more adept at capturing non-linear relationships between features and labels, were evaluated on their rank correlation results. The RF model demonstrated variability in terms of its accuracy in predicting the top 30 manager ranks, with 2014–2016 being slightly positive and more than doubling over 2017–2019. Figure 5 shows that during the same periods, the top decile consistently outperformed the bottom decile and further outperformed the category average when correlations increased. However, the RF model's performance was negatively affected by the near-zero and negative correlations observed in 2011 and 2021 for the top 30 managers and its poor accuracy in predicting the bottom 30 managers' ranks, which resulted in extremely negative rank correlations. This can be attributed to the inferior performance of the top decile relative to both the bottom decile and category average.

Figure 19 presents the rank correlation results of the XG model, which are consistent with the performance results shown in Figure 6. The XG model appears to have been most effective in identifying future top-performing managers during the 2011–2015 period, as top 30 rank correlations were mostly positive and exceeded 20%. The bottom 30 rank correlations were generally lower, with only 2014 and 2018 exceeding 30%. However, during the 2012–2015 period when correlations were positive, the top decile consistently outperformed the bottom decile by a significant margin, and this performance persisted until 2021.

The rank correlation results of the LG model, presented in Figure 20, shed light on its superior performance over the sample period. The LG model's accuracy is evident in terms of top 30 rank correlation, with seven years out of the eleven-year period showing positive correlations, five of which are more than 15%, and two are greater than 30% (2015 and 2017). Similarly, the bottom 30 rank correlation is also positive in seven of the eleven-year period, with the largest underperformance of the bottom decile portfolio being in 2015 and 2016, both of which have the highest rank correlation (more than 30%). The extreme period of outperformance observed in Figure 7 over 2015–2017 can be largely explained by the large positive rank correlations on the top and bottom 30 managers. Importantly, even though rank correlation explains accuracy or "hit-rate," 2014 shows that, even in the presence of positive rank correlation, the top decile can underperform its bottom counterpart. However, the higher correlation on the bottom 30 leads to the bottom decile underperforming the category average.

5. Discussion and conclusion

The current literature on active manager performance in developed markets exhibits two key consistencies. First, the majority of active fund managers tend to underperform their benchmarks, particularly when evaluating performance on a net-of-fees basis. Second, a subset minority of managers demonstrate skill, whether it be through market-timing or stock selection ability, in nearly every manager universe assessed. Thus far, the literature has primarily focused on identifying generalised factors that explain manager performance, such as size, costs, fund flows and manager experience, while concentrating on returns explained on an ex-post basis. Building on the recent research by Wu *et al.* (2021), Kaniel *et al.* (2022) and DeMiguel *et al.* (2021), we examine the possibility of identifying skilled emerging market active equity managers using ML on an *ex ante* basis.

Our study is centred on the South African active equity managers classified under the ASISA South Africa General Equity category, sourced from MorningStar, covering the period from January 2002 to December 2021. Like Wu *et al.* (2021), our study concentrates on performance-based predictors and utilises 36 features across six ML models that map these predictors to future manager excess performance. The models are trained on historical data and are applied at each portfolio formation date, where managers are sorted into equally weighted decile portfolios based on their probability of outperformance. The portfolio returns are then computed for the subsequent 12-month period, after which the process is repeated. Hence, all portfolio formation information, including manager probability and ranking, is generated in-sample and then applied out-of-sample to generate decile portfolio performance.

The present study investigates the applicability of ML models in predicting the performance of active fund managers in emerging markets. Our findings are consistent with the developed market literature, which suggests that ML models that can capture non-linearities are more effective in generating out-of-sample performance. Specifically, we find that the gradient boosting decision tree models, LG and XG, outperform other models in terms of nominal long-only and long-short excess returns. LG achieves an average annual return of 10.5%, while XG achieves 9.5%, both of which exceed the category average of 9.3% over the same period. Additionally, both models perform well in identifying future underperforming managers, achieving excess average long-short returns of 3.3 and 0.9% per annum, respectively. However, on a risk-adjusted basis, only the LG model produces statistically significant long-only alphas of 0.14 and 0.17% per month, translating to annualised alphas of 1.7 and 2% per annum. The annualised alphas are around 1.5% lower than those of similar strategies applied on US active managers as reported by DeMiguel *et al.* (2021).

To provide a benchmark for the effectiveness of the ML models, we also assess the performance of naïve manager selection methods, such as long-term reversal, momentum and

style alpha. We find that from a long-only perspective, the naïve top decile portfolios perform better than the ML top deciles on aggregate. However, ex post decile portfolio performance generated via LG is still superior on both a long-only and long-short basis.

Given that equally weighted cumulative and discrete period returns can conceal ML model precision, we further analyse the accuracy of the models using two tests. The first test focuses on long-only prediction success and the second on long (top 30) and short (bottom 30) rank correlation. For the former, we examine the top five to twenty managers per model in each sorting period and find that the RF model maintains the highest level of prediction precision, oscillating from 55.2% when considering the top five managers to 53% when considering the top twenty. Interestingly, LG only shows significant prediction accuracy when considering from the top fifteen managers onwards. Our second prediction accuracy test uses rank correlation analysis between the respective ML predicted *ex ante* rank and ex-post manager performance rank. The results show a high level of consistency with discrete period return performance and clearly indicate that the source of LG and XG's returns is attributable to ex ante ranking ability.

In this study, we present unique insights on the application of ML techniques for selecting active managers in emerging markets. Our findings contribute to the existing global and emerging market literature by identifying the superior performance of ML models with higher non-linear capabilities, particularly LG. However, our results also indicate that ML-based strategies are less profitable on aggregate when applied to emerging market asset managers, which is contrary to the evidence found in developed markets. We observe that linear ML models perform poorly in identifying skilled managers and naive manager selection methods such as style alpha and momentum achieve superior out-of-sample returns. Only XG, beyond LG, produces an economically meaningful long-short excess return but is statistically insignificant. We suggest that this evidence could be due to several factors, including the limited liquidity and universe of assets in emerging markets and our shorter and more recent timeframe in comparison to the recent developed market studies. Our study provides a clear explanation of ML performance through the analysis of rank correlations, which demonstrate the relatively high level of forecast accuracy of both LG and XG models, explaining their long-only and long-short performance. Despite the evidence of lower nominal and risk-adjusted profits, our study confirms the benefit of ML-based manager selection and offers novel evidence supporting the application of non-linear ML models for selecting active equity managers in emerging markets. However, a significant limitation of our study is its sole focus on a single emerging market. Therefore, future research would benefit from increasing the dataset to include additional geographies across emerging markets, particularly those operating within economies that have well-developed and active institutional and retail investor bases.

Notes

1. Free float of Naspers has reduced dramatically due to the unbundling of Prosus (PRX) in 2019.
2. Generally referred to as regression via classification ("RvC" hereafter) (see Torgo and Gama, 1997; Janssen and Furnkranz, 2011).
3. The Sharpe Ratio is a popular metric used to evaluate the risk-adjusted performance of a return series and represents the ratio of gross returns net of the risk-free proxy scaled by the standard deviation of the excess return series. Sharpe ratios are statistically evaluated applying the methodology of Bailey and Lopez Del Prado (2012). The statistical test assumes that Sharpe ratios follow a normal distribution but are impacted by the number of observations, skewness and kurtosis of return series. The test statistic therefore accounts for all the above by adjusting the test statistic upwards based on the number of observations and downwards for skewness and kurtosis. Information Ratios are calculated using portfolio returns net of a market proxy to the standard deviation of these excess returns, generally referred to as tracking error.

References

Bailey, D.H. and Lopez de Prado, M. (2012), "The Sharpe ratio efficient Frontier", *Journal of Risk*, Vol. 15 No. 2, pp. 13-59, doi: 10.21314/jor.2012.255.

Ban, G., El Karoui, N. and Lim, A. (2018), "Machine learning and portfolio optimisation", *Management Science*, Vol. 64 No. 3, pp. 1136-1154, doi: 10.1287/mnsc.2016.2644.

Berk, J. and Green, R. (2004), "Mutual fund flows and performance in rational markets", *Journal of Political Economy*, Vol. 112 No. 6, pp. 1269-1295, doi: 10.1086/424739.

Berk, J.B. and Van Binsbergen, J.H. (2015), "Measuring skill in the mutual fund industry", *Journal of Financial Economics*, Vol. 118 No. 1, pp. 1-20, doi: 10.1016/j.jfineco.2015.05.002.

Blake, C. and Morey, M. (2000), "Morningstar ratings and mutual fund performance", *Journal of Financial and Quantitative Analysis*, Vol. 35 No. 3, pp. 451-483, doi: 10.2307/2676213.

Brands, S., Brown, S. and Gallagher, D. (2005), "Portfolio concentration and investment manager performance", *International Review of Finance*, Vol. 5 Nos 3-4, pp. 149-174, doi: 10.1111/j.1468-2443.2006.00054.x.

Breiman, L. (2011), "Random forests", *Machine Learning*, Vol. 45 No. 1, pp. 5-32, doi: 10.1023/a:1010933404324.

Carhart, M. (1997), "On persistence in mutual fund performance", *The Journal of Finance*, Vol. 52 No. 1, pp. 57-82, doi: 10.1111/j.1540-6261.1997.tb03808.x.

Chang, C.E., Krueger, T.M. and Witte, H.D. (2022), "Enhancing equity index mutual fund returns using cost and morningstar ratings information", *The Journal of Investing*, Vol. 31 No. 3, pp. 122-141.

Chen, J., Hong, H., Huang, M. and Kubik, J. (2004), "Does fund size erode mutual fund performance? The role of liquidity and organization", *American Economic Review*, Vol. 94 No. 5, pp. 1276-1302, doi: 10.1257/0002828043052277.

Cortes, C. and Vapnik, V. (1995), "Support-vector networks", *Machine Learning*, Vol. 20 No. 3, pp. 273-297, doi: 10.1007/bf00994018.

Coulombe, P., Leroux, M., Stevanovic, D. and Surprenant, S. (2022), "How is machine learning useful for macroeconomic forecasting?", *Journal of Applied Econometrics*, Vol. 37 No. 5, pp. 920-964.

De Bondt, W. and Thaler, R. (1985), "Does the stock market overreact?", *The Journal of Finance*, Vol. 40 No. 3, pp. 793-805, doi: 10.1111/j.1540-6261.1985.tb05004.x.

DeMiguel, V., Gil-Bazo, J., Nogales, F. and Santos, A. (2021), "Can machine learning help to select portfolios of mutual funds?", *Barcelona GSE Working Paper Series*, Vol. 1245, pp. 1-40.

Fama, E.F. and French, K.R. (2010), "Luck versus skill in the cross-section of mutual fund returns", *The Journal of Finance*, Vol. 65 No. 5, pp. 1915-1947, doi: 10.1111/j.1540-6261.2010.01598.x.

Friedman, J.H. (2002), "Stochastic gradient boosting", *Computational Statistics and Data Analysis*, Vol. 38 No. 4, pp. 367-378, doi: 10.1016/s0167-9473(01)00065-2.

Fu, X., Du, J., Guo, Y., Liu, M., Dong, T. and Duan, X. (2018), "A machine learning framework for stock selection", *arXiv, arXiv:1806*, 01743.

Gallagher, D. (2003), "Investment manager characteristics, strategy, top management changes and fund performance", *Accounting and Finance*, Vol. 43 No. 3, pp. 283-309, doi: 10.1111/j.1467-629x.2003.00092.x.

Gil-Bazo, J. and Ruiz-Verdú, P. (2009), "The relation between price and performance in the mutual fund industry", *The Journal of Finance*, Vol. 64 No. 5, pp. 2153-2183, doi: 10.1111/j.1540-6261.2009.01497.x.

Gittelsohn (2019), available at: https://www.bloomberg.com/news/articles/2019-09-11/passive-u-s-equity-funds-eclipse-active-in-epic-industry-shift

Grinblatt, M. and Titman, S. (1994), "A study of monthly mutual fund returns and performance evaluation techniques", *Journal of Financial and Quantitative Analysis*, Vol. 29 No. 3, pp. 419-444, doi: 10.2307/2331338.

Janssen, F. and Furnkranz, J. (2011), "Heuristic rule-based regression via dynamic reduction to classification", *Twenty-Second International Joint Conference on Artificial Intelligence*.

Jegadeesh, N. and Titman, S. (1993), "Returns to buying winners and selling losers: implications for stock market efficiency", *The Journal of Finance*, Vol. 48 No. 1, pp. 65-91, doi: 10.1111/j.1540-6261.1993.tb04702.x.

Kacperczyk, M., Sialm, C. and Zheng, L. (2005), "On the industry concentration of actively managed equity mutual funds", *The Journal of Finance*, Vol. 60 No. 4, pp. 1983-2011, doi: 10.1111/j.1540-6261.2005.00785.x.

Kacperczyk, M., Nieuwerburgh, S. and Veldkamp, L. (2014), "Time-varying fund manager skill", *The Journal of Finance*, Vol. 69 No. 4, pp. 1455-1484, doi: 10.1111/jofi.12084.

Kaniel, R., Lin, Z., Pelger, M. and Van Nieuwerburgh, S. (2022), "Machine-learning the skill of mutual fund managers", *National Bureau of Economic Research*, w29723, pp. 1-57.

Krollner, B., Vanstone, B. and Finnie, G. (2010), "Financial time series forecasting with machine learning techniques: a survey", *Proceedings of the 18th European Symposium on Artificial Neural Networks, ESANN 2010*, pp. 25-30.

Li, B. and Rossi, A. (2020), "Selecting mutual funds from the stocks they hold: a machine learning approach", *SSRN*, 3737667, pp. 1-55.

Ludwig, R.S. and Piovoso, M.J. (2005), "A comparison of machine-learning classifiers for selecting money managers", *Intelligent Systems in Accounting, Finance and Management: International Journal*, Vol. 13 No. 3, pp. 151-164, doi: 10.1002/isaf.262.

Page, D., McClelland, D. and Auret, C. (2022), "Style rotation on the JSE", *Finance Research Letters*, Vol. 46, 102504, doi: 10.1016/j.frl.2021.102504.

Rasekhschaffe, K. and Jones, R. (2019), "Machine learning for stock selection", *Financial Analysts Journal*, Vol. 75 No. 3, pp. 70-88, doi: 10.1080/0015198x.2019.1596678.

SPIVA (2021), "Spglobal", available at: https://www.spglobal.com/spdji/en/documents/spiva/spiva-us-year-end-2021.pdf

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58 No. 1, pp. 267-288, doi: 10.1111/j.2517-6161.1996.tb02080.x.

Torgo, L. and Gama, J. (1997), "Regression using classification algorithms", *Intelligent Data Analysis*, Vol. 1 No. 4, pp. 275-292, doi: 10.3233/ida-1997-1405.

Wu, W., Chen, J., Yang, Z. and Tindall, M. (2021), "A cross-sectional machine learning approach for hedge fund return prediction and selection", *Management Science*, Vol. 67 No. 7, pp. 4577-4601, doi: 10.1287/mnsc.2020.3696.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 67 No. 2, pp. 301-320, doi: 10.1111/j.1467-9868.2005.00503.x.

Appendix

The 27four quantitative research team conducts style-based analysis on the top 100 shares (by market capitalisation) listed on the Johannesburg Stock Exchange ("JSE" hereafter). The research conducted involves style back-testing on the full cross-section of shares listed on the JSE over the period January 2000 to December 2021. For each style, shares are sorted on several style proxies. Long-only style benchmark returns are calculated on a buy-and-hold basis applying market capitalisation and factor weighting. Market capitalisation weighting implies assigning each share an initial in-portfolio weight based on their sorting date market capitalisation while factor weighting implies applying the standardised style score (proxy) as the initial weighting. Importantly, both market capitalisation and factor weighted portfolios are identical in terms of their share constituents, the only difference being the initial weight assigned at each portfolio sort date.

The proxy for each style is described on the table below:

Style	Proxy	Style factor name/Detail	Inverse style factor
Momentum	Historical cumulative return measured over the previous 252 trading days excluding the most recent 21 days of return	Winner = top 30 shares based on momentum rank	Loser = bottom 30 shares based on momentum rank
Value	Linear combination of book-to-market ratio (B/M), earnings yield (EY) and cashflow per share (CFPS)	Value = top 30 shares based on z-score combination of B/M, EY and CFPS	Growth = bottom 30 shares based on z-score combination of B/M, EY and CFPS
Low volatility	Linear combination of idiosyncratic risk and market beta measured over the previous 252 trading days	Low Vol = bottom 30 shares based on z-score combination of standard deviation and market beta	High Vol = top 30 shares based on z-score combination of standard deviation and market beta
Quality	Linear combination of return on assets (ROAs), return on equity (ROE) and debt to equity ratio (D/E)	Quality = top 30 shares based on z-score combination of ROA, ROE less D/E	Junk = bottom 30 shares based on z-score combination of ROA, ROE less D/E
Size	Natural logarithm of market capitalisation on portfolio sort date	Small = bottom 30 shares based on natural log of market capitalisation	Big = top 30 shares based on natural log of market capitalisation
Liquidity	Average value traded measured over the previous 252 trading days relative to market capitalisation	Illiquid = bottom 30 shares based on value traded scaled by market capitalisation	Liquid = top 30 shares based on value traded scaled by market capitalisation
Dividend yield	One year trailing total dividend (inclusive of special dividends) scaled by the current share price	High Yield = top 30 dividend paying shares based on natural log of dividend yield	Low Yield = bottom 30 dividing paying shares based on natural log of dividend yield
Currency risk	Currency beta measured over the previous 252 trading days	Rand hedge = top 30 shares based on currency beta	Rand tracker = bottom 30 shares based on currency beta

Source(s): Table by authors

Table A1. Style proxy and anti-factor definition applied to the estimation of time-series style returns

Lasso	{“penalty”:[“l1”],“solver”:[“liblinear”,“saga”],“C”:[0.01,1,10,100],“max_iter”:[100,1000,10000],“fit_intercept”:[True, False], “multi_class”:[“auto”, “ovr”, “multinomial”]}
Elastic net	{“loss”:[“modified_huber”, “squared_hinge”],“penalty”:[“elasticnet”], “alpha”:[0.01,1,10,100],“l1_ratio”:[0.05,0.15,0.35],“fit_intercept”:[True, False], “max_iter”:[100,1000,10000]}
SVM	{“C”: [0.01,1,10,100],“kernel”:[“linear”, “poly”, “rbf”, “sigmoid”],“gamma”:[“scale”, “auto”]}
Random forest	{“n_estimators”:[10,20,50,100],“criterion”:[“gini”, “entropy”],“max_features”:[“auto”, “sqrt”, “log2”],“min_samples_leaf”:[1, 2, 4, 8, 12, 24],“bootstrap”:[True, False]}
LightGBM	{“boosting_type”:[“gbdt”,“dart”,“goss”],“num_leaves”:[8,16,31], “learning_rate”:[0.01,0.05,0.1,0.3],“n_estimators”:[10,20,50,100]}
XGBoost	{“n_estimators”:[10,20,50,100],“max_depth”:[3,6,10,13,16],“booster”:[“gbtree”,“gblinear”,“dart”],“learning_rate”:[0.01,0.05,0.1,0.3]}

Note(s): Hyper-parameter tuning is conducted on 70% of the training data and evaluated on 30% of training data that is reserved for testing

Source(s): Table by authors

Table A2. Description of hyper-parameter grid dictionaries that are applied at each portfolio sort assuming $k = 5$ folds

Table A3.
This table describes the training data window as well as the train and test accuracy scores (number of correct classifications divided by the total number of labels) for each hyper-parameter tuned model

	Lasso		Elastic net		SVM		Random forest		LightGBM		XGBoost	
Training data	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
2002–2009	62%	33%	68%	20%	92%	4%	100%	13%	100%	19%	100%	27%
2002–2010	0%	0%	52%	31%	0%	0%	100%	20%	100%	43%	18%	0%
2002–2011	0%	0%	42%	28%	0%	0%	98%	47%	98%	47%	100%	38%
2002–2012	100%	78%	71%	86%	0%	0%	71%	60%	100%	100%	65%	86%
2002–2013	75%	12%	60%	39%	0%	0%	99%	45%	96%	23%	97%	29%
2002–2014	100%	66%	60%	50%	0%	0%	93%	77%	64%	71%	67%	46%
2002–2015	0%	0%	0%	0%	0%	0%	99%	67%	100%	25%	21%	0%
2002–2016	75%	97%	50%	50%	50%	50%	100%	34%	15%	0%	61%	25%
2002–2017	75%	56%	100%	91%	66%	51%	100%	53%	98%	61%	69%	34%
2002–2018	71%	60%	82%	64%	78%	7%	95%	47%	95%	70%	63%	52%
2002–2019	62%	54%	77%	4%	69%	51%	82%	56%	94%	66%	61%	59%

Note(s): A number of instances indicate model misspecification i.e. when both train and test accuracy are zero as well as several instances of overfitting where training accuracy is significantly larger than test accuracy. Importantly, the focus of the study is less concerned with the definition of the optimal set of hyper-parameters but rather more focused on providing each model with an "equal" tuning opportunity
Source(s): Table by authors

Prediction period (Yt +1)	Feature Period (Xt)	Lasso	Elastic net	SVM	Random forest	LightGBM	XGBoost
2011	2006–2010	50%	39%	43%	30%	38%	42%
2012	2007–2011	0%	12%	0%	16%	34%	5%
2013	2008–2012	0%	51%	0%	58%	15%	67%
2014	2009–2013	0%	0%	0%	23%	9%	2%
2015	2010–2014	0%	64%	0%	50%	100%	67%
2016	2011–2015	50%	35%	0%	50%	60%	44%
2017	2012–2016	0%	0%	0%	25%	21%	1%
2018	2013–2017	0%	0%	42%	0%	11%	50%
2019	2014–2018	50%	0%	69%	57%	57%	57%
2020	2015–2019	59%	54%	70%	53%	46%	59%
2021	2016–2020	39%	34%	48%	31%	31%	46%

Note(s): The first column describes the label/prediction period while the second details the most recent feature window. The results in the remaining columns detail the predictive accuracy of model (precision) when deployed on out-of-sample data. As described in the study, each model was used to forecast future probability of outperforming the market, therefore a low accuracy score need not necessarily entail a poor performing model but rather the total number of correct classifications relative to the total number of labels

Source(s): Table by authors

Table A4. This table details out-of-sample accuracy based on optimised models described per Table A3

Feature name		Lasso	Elastic net	SVM	Random forest	LightGBM	XGBoost
Cumulative return	1 Year	7	*5*	7	*1*	*1*	32
	2 Year	11	26	11	7	*3*	14
	3 Year	25	12	27	*4*	9	30
	4 Year	8	28	6	6	16	*4*
	5 Year	33	36	35	*2*	*5*	*1*
Style alpha		*5*	*2*	*2*	9	21	36
Information ratio (benchmark)	1 Year	30	35	34	*5*	*4*	*2*
	2 Year	*1*	7	*3*	13	14	27
	3 Year	21	16	28	24	11	13
	4 Year	35	16	31	17	24	12
	5 Year	6	*3*	*4*	20	8	33
Information ratio (Peer Avg.)	1 Year	27	16	24	22	12	17
	2 Year	32	16	21	28	16	8
	3 Year	19	16	*5*	21	25	22
	4 Year	*3*	16	14	27	21	15
	5 Year	26	9	23	18	21	26
Information ratio (style proxy)	1 Year	29	30	33	12	6	7
	2 Year	31	29	29	16	*2*	10
	3 Year	*4*	15	12	*3*	13	21
	4 Year	34	27	32	19	10	16
	5 Year	20	8	20	15	26	25
Information RATIO (style peers)	1 Year	*2*	*1*	*1*	11	19	34
	2 Year	36	34	36	25	7	*5*
	3 Year	17	16	17	14	14	20
	4 Year	23	16	18	10	16	19
	5 Year	10	14	15	8	20	29
Consistency vs bench		18	6	22	34	34	31
Consistency vs peer bench		28	32	30	29	35	9
Consistency vs style bench		14	16	25	30	31	18
Consistency vs style peer bench		12	*4*	10	31	27	35
Weighted consistency (bench.)		9	13	9	33	30	23
Weighted consistency (Peer bench.)		13	16	8	32	29	11
Weighted Consistency (Style Bench.)		22	31	13	26	33	6
Weighted consistency (style peer bench.)		15	11	16	23	32	28
Best performing style		24	33	26	34	28	*3*
Small/illiquid holding		16	10	19	34	36	24

Table A5.
This table describes the aggregate feature importance rank across the ML models applied

Note(s): The average top five ranks are emboldened in green font and indicates consistency when grouping models based on their adeptness to identify non-linearities. As an example, both style alpha and 1 year information ratio based on similar style peers achieves high average ranks for the linear models. Conversely, five year cumulative return and 1 year information ratio based on the category average achieve the highest average ranks across the non-linear models
Source(s): Table by authors

Corresponding author
Yudhvir Seetharam can be contacted at: yudhvir.seetharam@wits.ac.za