Neural Computing and Applications (2024) 36:11963–11986
https://doi.org/10.1007/s00521-024-09700-3

ORIGINAL ARTICLE

Machine learning-based stock picking using value investing and quality features

Ronen Priel[1,2] · Lior Rokach[1]

Received: 1 June 2023 / Accepted: 25 March 2024 / Published online: 18 April 2024

Abstract

Value Investing stands as one of the most time-honored strategies for long-term equity investment in financial markets, specifically in the domain of stocks. The essence of this approach lies in the estimation of a company's "intrinsic value," which serves as an investor's most refined gage of the company's true worth. Once the investor arrives at an estimation of the intrinsic value for a given company, she proceeds to contemplate purchasing the company's stocks solely if the prevailing market price of the stocks significantly deviates below the estimated intrinsic value, thus presenting an enticing buying opportunity. This deviation, referred to as the "margin of safety," represents the disparity between the intrinsic value and the current market capitalization of the company. Within the scope of this endeavor, our objective is to automate the stock selection process for value investing across a vast spectrum of US companies. To accomplish this, we harness a combination of value-investing principles and quality features derived from historical financial reports and market capitalization data, thereby enabling the identification of favorable value-driven opportunities. Our methodology entails the utilization of an ensemble of classifiers, where the class is determined as a function of the margin of safety. Consequently, the model is trained to discern stocks that exhibit value characteristics warranting investment. Remarkably, our model attains a success rate surpassing 80%, effectively identifying stocks capable of yielding an annualized return of 15% within a three-year timeframe from the recommended stock purchase date provided by the model.

Keywords Stock picking · Value investing · Quality investing · Random forest · Gradient boosting trees

1 Introduction

Publicly traded companies listed on stock exchanges, such as NASDAQ and NYSE in the USA, are obliged to periodically disclose their operational and financial results to the wider investment community. This transparency enables both existing and potential shareholders to access the necessary data for making well-informed investment decisions. These disclosures, known as the company's Fundamentals, follow the structure and content prescribed by national and international accounting standards, primarily IFRS [1] and US-GAAP [2]. Fundamentals comprise three key components: the Income Statement, Balance Sheet, and Cashflow Statement. Our research centers around fundamental analysis for investment purposes, with a specific focus on automating a prominent fundamentals-based investment approach, namely Value Investing.

Value Investing entails a strategic investment methodology that involves gaging the intrinsic value of a company, representing an investor's most discerning assessment of the company's true worth. Subsequently, stocks that exhibit a substantial downward deviation from their intrinsic value are selectively purchased, providing an opportunity to capitalize on their undervaluation. At the heart of value investing lies the recognition that markets are not entirely efficient, meaning that in the short-term stocks may be either overpriced or underpriced relative to their intrinsic value. However, value investors firmly

✉ Ronen Priel
ronenpr@post.bgu.ac.il

Lior Rokach
liorrk@post.bgu.ac.il

[1] Present Address: Ben-Gurion University of the Negev, Beer Sheva, Israel

[2] Tel Aviv, Israel

believe that, over the long run, financial markets tend to be predominantly efficient, causing underpriced stocks to eventually converge toward their intrinsic value within a matter of months or years. To ascertain a company's intrinsic value, value investors employ a range of valuation methods. These methods typically involve an analysis of the company's historical fundamentals, combined with an assessment of its present and anticipated future conditions encompassing factors such as growth rates, profitability, competition intensity, and management quality. Ultimately, these analyses lead to a final numerical estimate or range representing the intrinsic value.

The fundamental principle underpinning value investing is the concept of the margin of safety (see Fig. 1). A value investor will only acquire a stock if it is significantly undervalued, often trading at a substantial discount (e.g., 20–70%) below its estimated intrinsic value. The margin of safety serves as a risk management mechanism for the investor, mitigating the impact of inevitable valuation errors inherent in a subjective and multifaceted process.

In our research, our objective is to automate the stock selection process in the style of value investing. We diligently monitor thousands of stocks on a weekly basis, assessing whether a trained machine learning model can effectively identify value investing opportunities at the right time, as measured by the resultant returns derived from these investments.

This paper proceeds as follows. Section 2 describes the relevant literature and highlights the novelties of our paper. Section 3 explains in detail the methodology used. Section 4 describes the results of our experiments, while Sect. 5 includes a deeper analysis of these results, as well as features importance analysis, a construction of an example portfolio based on the model's stock picks, and a list of future work items.

2 Related work

2.1 Value investing

Value investing, a structured investment methodology introduced by Graham & Dodd [3], centers around investing with a margin of safety below intrinsic value. We have drawn upon several authoritative sources [4–13] to summarize the key characteristics of this methodology.

2.1.1 Stocks mispricing

Value investors acknowledge that the market exhibits partial efficiency, leading to frequent short-term mispricing of stocks. However, over the long term, stocks are expected to converge toward their intrinsic value. Value investors view this volatility as an opportunity to purchase stocks at significant discounts. The reasons for stock mispricing include wide market optimism or pessimism, supply and demand disruptions caused by institutional investors, market overreactions to news, and the undervaluation of "under the radar" companies, particularly smaller-cap stocks.

2.1.2 Risk awareness and loss avoidance

The detrimental impact of losses on a portfolio's long-term returns is well-recognized. A negative return of X% necessitates a significantly higher positive return than X% to recover the loss. An illuminating example can be found in [5], where an investor who earns 16% annual returns over a decade accumulates more wealth than an investor who earns 20% annual returns for nine years and then experiences a 15% loss in the tenth year. Consequently, value investing places paramount importance on risk, defined as the expected loss associated with an investment in a stock. This underlies the concept of the margin of safety: when a company's intrinsic value is relatively accurately estimated and its stock trades at a significant discount, the long-term probability of loss is low, while the likelihood of substantial long-term returns is high.

2.1.3 Contrarian look

Value investors seek out stocks that are substantially undervalued compared to their estimated intrinsic value. These stocks are typically overlooked or sold by most other investors. As articulated by a renowned value investor, "Buy [stock] at the point of maximum pessimism [in the market regarding this stock]" [14]. The same contrarian approach applies to selling stocks. This necessitates a

Fig. 1 Value investing with margin of safety

particular character on the part of value investors, including unwavering belief in their judgment, patience, and resistance to peer pressure. Many successful value investors consider the psychological challenge of going against the crowd to be the most arduous aspect of value investing.

2.1.4 Bottom-up investing

Value investing adopts a bottom-up approach, where stocks are meticulously selected based on detailed analysis of their unique current financial and operational state. This distinguishes value investing from top-down approaches such as factor investing, even if the factor used involves a value metric.

2.1.5 Valuation

Given that the value of any financial asset represents the risk-adjusted discounted future cash flows, achieving a highly accurate point estimation is typically challenging. A common alternative approach involves considering multiple future scenarios with varying probabilities and derived intrinsic values. These scenarios are then aggregated using a weighting mechanism to arrive at a final range estimation of intrinsic value. Various valuation methods are employed in practice, including Discounted Cash Flows (DCF) and Multiples [15, 16], valuation based on assets and ongoing business [13], and liquidation value-based methods. Regardless of the specific method used, value investors prioritize verifying that the margin of safety is sufficiently large by establishing a more certain lower bound of valuation. Investment decisions based on this lower bound significantly reduce the risk of loss.

2.2 Quality investing

The Finance academic community has dedicated decades to investigating factors, the drivers of equity market returns, as evidenced by numerous studies [17–21]. Harvey et al. [22] revealed that a staggering 316 factors have already been analyzed and published, and at the current rate of discovery, we can expect approximately 600 factors to be identified by 2035. These factors span various domains such as macroeconomics (e.g., exposure to inflation), sectors, geographical areas, investment styles, and more. Among these factors, both value and quality hold significant prominence.

The financial industry employs various definitions of "quality," all of which rely on a company's fundamentals. Most definitions revolve around several recurring themes:

- Profitability and Growth: Measures the rates of profitability and growth exhibited by companies.
- Efficiency: Gages the effectiveness of a company in its investments and cash generation.
- Consistency: Evaluates the long-term stability of a company's performance.
- Leverage: Assesses a company's ability to fulfill its debt and other obligations without risk.

Both academic and industry evidence consistently suggest a positive correlation between the combination of value and quality factors and excess returns in the equity market. Novy-Marx [23] compares the performance, measured as the resulting excess return (alpha) over a well-established factor model, generated by seven standalone quality metrics and their integration with value factors. The analysis demonstrates that many of the utilized quality metrics contribute to an annual 4% + premium over the performance of the seminal FF-3 factor model [20], which already includes value as one of its three factors. Moreover, an integrated value-quality factor model outperforms a value-only strategy, as measured by alpha over the CAPM [24].

Additional evidence can be found in the 2021 annual shareholders letter of Fundsmith [25], a leading UK-based money management fund overseeing approximately $35 billion in assets under management (AUM). The fund follows a clear set of quality criteria in its investment approach, coupled with a value-based screening to avoid overpayment, which often poses a challenge when dealing with high-quality companies. Similarly, O'Shaughnessy Asset Management [26], an American factor investing firm managing around $5 billion in AUM, achieved double the return of a value-only factor portfolio by incorporating three factors—value, quality, and dividend yield—in their portfolio construction.

2.3 Past ML papers on fundamentals-based investing

In the realm of stock market prediction, a substantial amount of research has been focused on forecasting short-term movements using daily or sub-daily stock prices and trade volumes as features. However, the endeavor to predict long-term stock returns for specific stocks based on fundamental financial reporting has received comparatively less attention.

Earlier studies in this domain relied on relatively small datasets. Olson and Mossman [27] tried to predict one-year-ahead stock returns for approximately 2350 stocks using 61 accounting ratios as input. The authors concluded that neural networks yielded the best results, achieving an annual premium return of 10% over the market total return. Cao et al. [28] explored stock return prediction in the Chinese stock market and demonstrated that feedforward

neural networks (FFNN) outperformed linear models. Kryzanowski et al. [29] analyzed 150 companies over a five-year period in the 1990s and found that FFNN successfully classified 72% of positive/negative returns in a one-year-ahead period using financial ratios and macroeconomic variables.

More recent investigations have employed larger datasets. Abe and Nakayama [30] utilized a 26-year dataset encompassing 25 major fundamentals of approximately 320 stocks in the Japanese stock exchange. They predicted each stock's return one month ahead by analyzing the fundamentals of the past five quarters as input features. The results demonstrated that deep neural networks generally outperformed shallow neural networks, and the top-performing neural networks also outperformed Support Vector Machines (SVM) and Random Forest (RF).

Alberg and Lipton [31] employed 20 fundamentals and four momentum features from over 11,000 companies spanning 40 years of data to train a deep neural network. Their model forecasted future fundamental data based on a trailing five-year window. When compared to the FF-3 factor model [20], the resulting model yielded an additional compounded annual return of 2.7%.

Yang et al. [32] leveraged 27 years of fundamentals from 1142 companies included in the S&P 500 index during the period 1990–2017. They used 20 fundamentals as features and aimed to predict the return of each stock by training the model on an expanding training set spanning 16–40 quarters. Five different models, including Gradient Boosting Machine (GBM), were trained, and the best-performing model from the last quarters was used to predict stocks for the upcoming quarter. This strategy yielded an annualized return of around 11% compared to the S&P 500's approximate 5% in the out-of-sample period.

2.4 Innovation of our research

Our research focuses on creating a novel machine learning model to pick stocks with expected positive returns using fundamentals-based value investing concepts. As such, our main point of comparison is the fundamentals-based papers listed in Sect. 2.3. Our research introduces the following novel elements compared to these papers:

- Scale and Timely Actions: Our model allows for a more frequent buy/sell decisions per stock over a large (scale of thousands) universe of stocks—our model makes weekly buy/sell decisions vs. quarterly in [32], annual in [27, 29, 31], and monthly in [30].
- Superior Value Features: Previous papers [20, 27, 29–32] used the current level of a common financial multiple (such as Stock Price/Book Equity) to rank all the stocks in the universe to decimals, using the lower decile as "cheap" stocks and the higher decile as "expensive" stocks. We do not subscribe to such a definition, as it ignores the significant differences in the common levels of the multiple among different industries, driven from each industry's unique characteristics (intense of competition, profitability, demand growth, etc.). Instead, we calculate the value features by comparing each symbol only to other companies in its industry. In addition, we do not use deciles for defining "cheap" and "expensive" stocks, but instead compare a company's current multiple to the historical median of the industry, resulting in a more accurate value feature.
- Furthermore, while the above-mentioned papers utilize fixed multiples over all stocks through all the research period (which can last months or years), our model re-chooses every week the most appropriate multiple to use for each stock. This prevents our model from using multiples which are less suitable for a specific company.
- Class Definition: We employ an innovative function (class_type), used to calculate a binary class label for one stock at one date by examining both the level of cheapness of the stock at that date, and the return is achieved within the class's time horizon. This definition enables the model to locate stocks in the test set which are both cheap and expected to rise significantly within the class's time horizon. By using this approach, we aim to employ the stock-picking technique used manually by leading value investors [3–9] within the machine-learning academic community for the first time, as we could not find this definition or a similar one in any of the papers we reviewed [20, 27–32].

3 Methods

3.1 General concept

We employ an ensemble of binary classifiers to select stocks in a value investing style on a weekly basis from a vast universe with over 2000 stocks. The input to the model includes value features to assess the cheapness of a stock and quality features to gage a company's operational and financial prospects. The model goal is to correctly identify stocks that are significantly undervalued compared to their intrinsic value and have the potential to close this pricing gap within 1–3 years.

3.2 Raw financial data

Our primary data source consists of the historical fundamentals of 2161 companies, extracted from the three

components of quarterly financial reporting: the Income Statement, Balance Sheet, and Cash Flow Statement. We focused on data from the years 2000 to 2019, ensuring that each stock had at least 10 consecutive years of complete raw data (in other words, each of the companies has between 10 and 20 consecutive years of full data ending in end of 2019). This trading universe of 2161 companies represents a diverse range of stocks traded in the leading US stock exchanges (NYSE, AMEX, NASDAQ) at the end of 2019, both in terms of market capitalization size and industry sector. The universe does not include penny stocks (stocks with very low stock price which are a riskier investment tool). The data were obtained from [33]

Furthermore, we use the weekly market capitalization and the industry of each company, obtained from [34] and [35].

Although the raw data are subject to survivorship bias, we believe its size is sufficient to mitigate this bias and yield reliable results.

3.3 Features extraction

The model's features are calculated using raw fundamentals data. For clarity of reading, we did not include here the full definition of each finance/accounting term appearing in Sects. 3.3.1–3.3.3 below and instead collected these definitions into a Financial Terms Glossary Appendix at the end of the manuscript (see Appendix A).

3.3.1 Value features

We leverage the insights of Van Den Berg [7] to compute the value features for each company on a weekly basis. The value features are derived from the historical distribution of three value multiples: Price/Earnings, Price/Book, and Enterprise Value/EBITDA, specific to the company's industry classification under the GICS framework.

For each company each week, we choose the most relevant multiple to use by determining the correlation between the past industry distributions of these multiples and the company's past market capitalization distribution. This correlation is quantified as the percentage of weeks where both the company's market capitalization and the industry's multiple fell within the lowest α percentile. We compute this correlation for α values of 5%, 10%, 15%, and 20%. Among the twelve calculated correlations (3 multiples * 4 α values), we select the multiple that exhibits the highest correlation as the foundation for our value features (for this specific company at this specific week). We use the measured correlation and α as value features in our model (features 5–6 in Table 1).

Once a multiple was chosen for a company in a specific week, we compare the chosen multiple value v of the company in that week to the company's industry's past distribution D_i of the same chosen multiple. We use the percentile of v within D_i as one feature, and the ratio of v to the 25%, 50% and 75% percentiles of D_i as three additional features (features 1–4 in Table 1).

This results in 6 value features which enter the model, as presented in Table 1.

3.3.2 Quality features

In our approach, we incorporate a diverse set of quality features based on the comprehensive work of Noby-Marx [23], which collected the most commonly used and significant set of factors that contribute to a company's quality measurement, including profitability, growth, capital allocation efficiency, leverage, accounting standards compliance levels, performance consistency, and more. These features originate from either academic research [36–38] or from leading investors which specialize in quality investing [4, 6, 39].

The quality features are calculated using information from the past 1–5 years, we mark the number of past years data required to calculate each feature in the "Years Back" column of Table 2. The features are either Boolean or a rank among all companies in the universe within the given week (We sort all companies based on the value of the quality feature being ranked, giving rank 1 to the company with most positive value (for some features larger value is better, for other larger value is worst), rank 2 to the company with the 2nd most positive value, and so on).

In total, we use 22 quality features per stock per week, summarized in Table 2.

3.3.3 Industry ID feature

The last feature we use is an ID of the industry of the company, based on Standard and Poor's GICS industry classification scheme [35].

3.4 Adding classes

3.4.1 The class_type function

Our ensemble model is composed of multiple binary classifiers, each trained and tested using a supervised learning framework. The binary label values used by each binary classifier are not necessarily the same, instead they are calculated using the class_type function. The class_type function receives two input parameters—the margin of safety (*mos*) and the number of years ahead (*yh*)—and calculates the binary label values for the data samples (we explain this calculation in detail in the Sect. 3.4.2 following). We employ 18 different $<$ mos, yh $>$ input pairs,

Table 1 Value features used by the model

No.	Feature
1	The percentile of current week's chosen multiple value within the industry's past distribution of the chosen multiple
2–4	Ratio of current week's chosen multiple value to the values of the 25%, 50% and 75% percentile of industry's past distribution of the chosen multiple
5	The measured correlation % with market capitalization
6	The α value used

Table 2 List of quality features

	Quality framework	Domain	Feature (all derived from fundamentals line items)	Years back	Type
1	Graham's	Profitability	Earnings Per Share (EPS) growth in last 5 years	5	Boolean
2	G-Score [4]	Efficiency	Dividends or buybacks in all last 5 years	5	Boolean
3		Consistency	Positive Net Income (NI) in each of last 5 years	5	Boolean
4		Leverage	Current Ratio above a threshold	1	Boolean
5		Leverage	Total Assets (TA)/Long-Term Debt (LTD) is positive	1	Boolean
6	Grantham quality rank [39]	Profitability	Return on Equity (ROE) rank	1	Rank
7		Consistency	Low ROE volatility over last 5 years	5	Rank
8		Leverage	Total Assets / Book Value	1	Rank
9	Piotroski's	Profitability	Positive NI & CFO (Cash Flow from Operations)	1	Boolean
10	F-score [36]	Profitability	Positive YoY NI growth	2	Boolean
11		Efficiency	Positive YoY change in Asset Turnover	2	Boolean
12		Efficiency	Positive YoY change in Gross Margin (GM)	2	Boolean
13		Efficiency	Sloan Accruals [40] > 0	2	Boolean
14		Leverage	Positive YoY changes in Current Ratio	2	Boolean
15		Leverage	Positive YoY changes in leverage ratio	2	Boolean
16	Greenblatt magic formula [6]	Profitability	Earning Yield	1	Rank
17		Efficiency	Return On Invested Capital (ROIC)	1	Rank
18	Novy-Marx GM [37]	Profitability	GM / TA	1	Rank
19	Mohanram	Profitability	Return On Assets (ROA) is better than industry median	1	Boolean
20	G-Score [38]	Efficiency	CFO > NI	1	Boolean
21		Efficiency	Investment Intensity is bigger than industry median	1	Boolean
22		Consistency	NI & Revenues Growth variance in last 5 years is bigger than industry median	5	Boolean

each with different resulted binary label per data sample. These < mos, yh > input pairs are derived from the cartesian product of the sets mos_set = [0.2, 0.3, 0.4, 0.5, 0.6, 0.7] and yh_set = [1, 2, 3] sets.

Each of the binary classifiers in our ensemble uses only one <mos,yh> input pair to class_type as its source of binary label values (we explain how we construct the ensemble in detail in Sect. 3.6.2 following). This structure enables us to measure how the performance of the model changes as we increase *mos or yh*.

3.4.2 Calculating label values formula

The two input parameters of class_type function are margin of safety (*mos*) and the number of years ahead (*yh*) it takes for a stock to reach its intrinsic value relative to the current week's date.

Value investors typically assume market efficiency with occasional short-term mispricing, allowing us to approximate the intrinsic value by using the chosen multiple's historical median value over the relevant GICS industry

multiplied by the appropriate ratio denominator (book value, earnings, or EBITDA). Thus, the binary label value of a data sample—meaning the label value of one stock in a specific week—is calculated by the class_type function as follows:

- Calculate the current_distance as the ratio of the current data sample's chosen multiple (a value feature) to the industry's historical median value for that multiple.
- Calculate max_return (*yh*) as the stock's maximum return within *yh* years after the current week.
- Set the data sample's binary label value to 1 if both conditions are met: current_distance is greater than or equal to *mos*, and max_return (yh) is greater than or equal to *mos*. Otherwise, the data sample's label value is set to 0.

This definition aligns with the traditional principles of value investing, where a positive class indicates that the stock is undervalued and expected to recover to its intrinsic value within the defined period (as illustrated in Fig. 1, the buy zone). One important aspect derived from this definition is that a label value at a specific date is a function of what happened to the stock price in the following *yh* years and hence requires *yh* years of future data to be usable.

3.4.3 Annualized return (AR)

For any financial asset, if the expected total return after n years is r, the expected annualized return (AR) is calculated as $\mathrm{AR} = \sqrt[n]{1+r} - 1$. Hence, the AR per input parameters pair (mos, *yh*) to class_type is $\mathrm{AR} = \sqrt[yh]{1+\mathrm{mos}} - 1$.

3.5 Final dataset

Our final dataset consists of 29 features (6 value features, 22 quality features, 1 industry feature) calculated per each company per week. The following factors affect the final number of data samples (recall that each data sample is uniquely identified by a <company, week date> tuple):

- The original number of raw data years available per company. The minimum is 10 years, the maximum 20 years.
- The first 5 years of raw data per company are used for features extraction only (recall that the quality features are calculated based on 5 years of past data) and are not part of the final features dataset which enters the model.
- For a specific *yh* input parameter of class_type, we cannot calculate the label value for the last *yh* years of raw data as we do not have the required future data yet (recall that a label value of a stock at a given date is a function of what happened to the stock price in *yh* years following the given date). This implies that we obtain different final dataset sizes for different pairs of input parameters to the class_type function.
- While each data sample is uniquely identified by a <company, week date> tuple, this tuple is not part of the features and does not enter the model in train & test phases. In other words, the model is not learning anything related to a specific company nor to a specific period, as it is not aware of this information. Instead, the model aims to identify if a company is in the "buy-zone" of Figure-1 for any company at any time.

Table 3 presents the resulting final dataset sizes and imbalance ratios for each class_type input parameters pair, highlighting the imbalanced nature of the dataset with positive/negative labels ratios ranging from 4 to 22%.

3.6 Ensemble model

3.6.1 Choice of classification algorithms

We first considered several types of classification algorithms, including Random Forrest (RF), Gradient Boosting Machine (GBM), Deep Learning (DL) and Logistic Regression (LR). This preliminary analysis was performed at relatively early stage of the research, using a subset of the final dataset used in the paper's final model (we used earlier version of the value features, a subset of the final quality features set, and no industry features were used) and only 3 input parameters pairs to class_type function (out of the 18 input parameters pairs used by the final model). We tested these algorithms with 2 different values per several dominant hyperparameters (see Appendix B for detailed list of these hyperparameters) using simple time-based train-test split. This preliminary analysis resulted in consistent superior mean precision and recall results of RF and GBM versus LR and DL (see Table 4). We hence chose to focus further only on an ensemble of RF and GBM models. The following paragraphs describe these two algorithms.

Table 3 Dataset size and imbalance ratio per class

mos	*yh* Total data samples	1 851,219	2 769,056	3 689,009
0.2	% Positive label	16%	20%	22%
0.3		11%	15%	17%
0.4		8%	12%	13%
0.5		6%	9%	11%
0.6		5%	7%	9%
0.7		4%	6%	7%

Table 4 Mean precision and recall results in the preliminary analysis stage

Algorithm	Precision	Recall
Random Forrest	**67%**	**62%**
Gradient Boosting Machine	**71%**	**73%**
Deep Learning	55%	62%
Logistic Regression	62%	44%

A decision tree is a hierarchical model where each internal node performs a univariate test, with branches indicating test outcomes. Typically, the most crucial variable is positioned at the tree's root. In a classification tree, each leaf represents a predicted class. To predict a class for a new corporate, we traverse the tree from the root to one of the leaves based on test outcomes. A compact decision tree is prized for its ease of following decisions along its branches.

Training a decision tree model initiates at the root node with the entire training set. Each internal node divides the training set into two subsets based on a tested variable's value: one subset meeting a specific criterion (e.g., below a cutoff value) and the other containing the remaining cases. Variable selection and cut-off values aim to maximize classification performance.

Despite their advantages, decision trees have drawbacks such as high variance (small data changes lead to different trees) and limited predictive accuracy compared to complex models. To address these issues, we employed an ensemble of trees, also known as a decision forest consisting of multiple trees whose outputs are combined. We employed two popular decision forest algorithms: Random Forest [41] and Gradient-Boosted Machines [42] (using decision trees as the predictors). Both these algorithms excel in deriving models from tabular data already represented using meaningful features defined by domain experts, as is our research's case.

Random Forest comprises multiple decision trees, each trained independently without pruning. To introduce diversity, randomness is injected into the learning process through two mechanisms: (1) each tree is trained on a random bootstrap sample of cases, and (2) during tree splitting, a random subset of variables is considered, with the best variable selected among them.

The GBM algorithm trains a decision forest sequentially, with each tree fitted to the pseudo-residuals of preceding trees using a logistic loss cost function. This approach allows combining numerous shallow classification trees, particularly when a low learning rate is used (e.g., below 0.1).

3.6.2 Structure of the ensemble model

Per each of the 18 class_type input parameters pairs (defined in Sect. 3.4.1), we employ 8 binary classifiers which derive the binary label value for each data sample by executing the class_type function with that pair as input parameters. These 8 classifiers per pair differ by the classification algorithm used and its hyperparameters:

- 4 RF models, each with a different number of trees (20, 30, 100, 200).
- We noticed that tuning other hyperparameters does not affect performance, hence we fixed all RF classifiers to use a maximum tree depth of 10, 'gini' as the split criterion, minimum of two samples are required to split an internal node, and a leaf can hold one or more data samples. No class weights are used.
- 4 GBM models, each with a different number of trees (30, 50, 100, 500). We noticed that tuning other hyperparameters does not affect performance, hence we fixed all GBM classifiers to use a depth of 5, logistic loss cost function, learning rate of 0.03, maximum 31 leaves per tree, a leaf can hold one or more data samples, L2 regularization parameter (lambda) of 3. No class weights are used.

The result is an ensemble of 144 binary classifiers (8 classifiers per each of the 18 class_types input parameters pairs), each uniquely identified by the following 3-tuple: <class_type, algorithm (RF/GBM), number_of_trees> .

3.6.3 Training process

The full dataset, ranging from 2000 to 2019, was divided into train and test sets, based on specific date as cutoff. All the data before this cutoff date compose the train set and are used to train the binary classifiers, while all data following this cutoff date compose the test set which is used only for prediction by the trained classifiers. We used 3 different such divisions, to verify that the model's performance is consistent among them. The three cut-off dates are 31 Dec 2011, 31 Dec 2012 and 31 Dec 2014.

Our model aims to maximize the prediction's precision, as this is the most important metric in a practical investment universe of thousands of stocks. Maximizing the precision enables investors to manage a successful portfolio even with a very low recall, meaning with a small number of picked stocks. In fact, if we could produce a model with 100% precision (meaning every stock picked by this model will surely increase in its market value as predicted by the model's *mos and yh* parameters) we would only need one stock to invest in at each point in time. As such, we performed a linear search for finding the binary

classification threshold (a data sample is classified as positive only if the classifier's probability of being positive is equal to or greater than that threshold) that would maximize precision. We identified quickly that this happens in thresholds in the range of [0.80..0.99] with 0.01 intervals. This results in 20 different thresholds which we employed to each of the 144 binary classifiers trained in our research, enabling us to capture the resulted precision and recall values.

While training each binary classifier, we used time-based cross-validation with 5 folds. The first fold includes the earliest 20% of the train set, the 2nd fold the second-earliest 20% of the train set, and so on. The training includes 4 different train and test cycles, where we first train on the first fold and test on the 2nd fold, then train on the 1st and 2nd folds and test on the 3rd fold, and so on. Per each of the 144 binary classifiers, and per each of the 20 binary classification thresholds, we calculate the mean precision and recall over the 4 cross validation train-test folds.

In total, we get 2880 train precision and recall results (144 binary classifiers * 20 binary classification thresholds per binary classifier), each uniquely identified by the 4-tuple:

<class_type, algorithm (RF/GBM), number_of_trees, binary_classification_threshold>.

3.6.4 Testing process

We do not use all the trained binary classifiers in the test process over the test set, and instead aim to choose a subset of trained binary classifiers which will result in high precision with sufficient recall. Hence, we predict over the test set only with the trained binary classifiers which comply with the following criteria (see Fig. 2):

- Minimum cross-validation train precision of 70% across at least one of the 20 binary classification thresholds.
- Minimum expected annualized return (AR) of 10% (comparable to the long-term average annual return of the S&P 500 and Russell 3000, which is approximately 10%). Recall that a binary classifier's AR is derived from the input parameters used by class_type to calculate its label values.

All binary classifiers that satisfied these criteria were retrained on the full train set. For binary classifiers which met the criteria several times with different binary classification thresholds, we created a separate binary classifier instance for each such threshold and retrained it on the full train set.

During this retraining process of each binary classifier, we adjust the test set's beginning date based on the *yh* parameter used by class_type function to calculate the classifier's binary label values. The test set technically begins on January 1 of the year following the train-test cutoff year. However, since we used lookahead information during training (for label value calculation) with a lookahead of *yh* years, we removed the data samples from the first *yh* years of the test set. As a result, the test period starts *yh* years after the cutoff year.

We performed multiple prediction cycles over the test set per binary classifier and used the mean prediction results as our final test results, enabling us to calculate the precision and recall values over the test set per each of the test process binary classifiers. These results are used to evaluate the overall performance of the entire ensemble model at two levels:

- ML level: We measure the resulted precision and recall from the test process, as well as compare them to the training precision and recall results.
- Investment Grade precision: We assess if the chosen stocks are "investment grade" stocks, defined as stocks with an expected annualized return above 15%. This criterion is based on the maximum return within 1, 2, or 3 years, measured from the buy date predicted by the model. We calculate the percentage of investment-

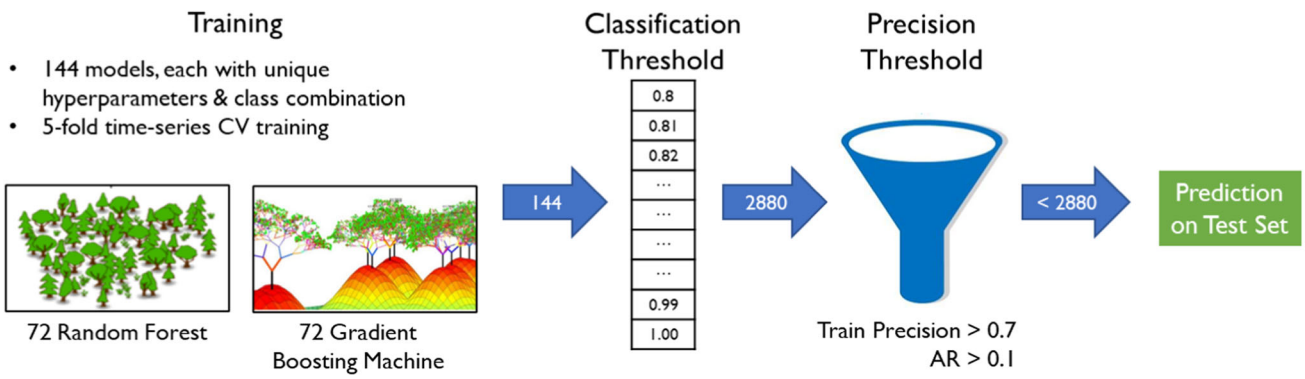

Fig. 2 Our Architecture of 144 models and 2880 precision results

grade stocks among all the stocks selected by the model ("investment grade precision").

Additionally, we construct a simple portfolio using these stocks and compare its performance against the Russell 3000 index (which is an appropriate benchmark as it includes both large and small companies, like our trading universe) over a 6-year test period.

4 Experiments results

4.1 Cross validation training results

The distributions of precision are comparable across the three cutoff years, with approximately 20% of the precision results exceeding 80%, and approximately 45–55% surpassing 70% (Fig. 3).

The distributions of recall are also comparable across the three cutoff years, with approximately 60–65% of the recall results is smaller than 10%, and approximately 20% between 10 and 30% (Fig. 4).

4.2 Test results: ML level

4.2.1 Precision–recall results

We conduct a comparison between the precision results of the test classifiers during the training and testing phases. Figure 5 illustrates the distribution of test precision across different cutoff years. It is evident that many test precision results fall within the range of 0.5–0.7, a decline in precision compared to the 0.7+ precision achieved by the same classifiers during the training stage. Statistical analysis confirms the significance of this deterioration, as demonstrated by a *T*-test over the 2012 cut-off year (t statistic = 2.163, p value = 0.0153).

More importantly, we checked the distribution of the test recall results for all the classifiers with train precision > 0.7 and for all the classifiers with both train and test precision $> 70\%$ (Tables 5 and 6, respectively). For example, in the first row of Table 5 we see that 41% of the 2880 classifiers (= 1172 classifiers) participated in the test process (as they achieved train precision higher than 70%), with 21% of the classifiers (604 classifiers) reach $< 10\%$ test recall and 20% (568 classifiers) reach $> 10\%$ test recall.

The results show that, as expected, there is a trade-off between precision and recall. Note the lower recall results of Table 6 which includes a smaller subset of classifier with higher test precision compared to Table 5. More importantly, for practical investment purposes these results are encouraging, as our model includes hundreds of classifiers with high test precision ($> 70\%$) while still producing a test recall bigger than 10%. Using these classifiers should result in a sufficient number of stocks picks, all with a high chance of resulting in high returns. A portfolio build by these stocks is likely to surpass in performance common benchmark indices like Russel-3000 or S&P-500.

4.2.2 Precision results analysis

To further investigate the deterioration in test precision vs the train precision, we examine its scale in relation to different decimal values of the train precision above the 0.7 threshold. Table 7 reveals that the deterioration becomes more pronounced for higher train precision values. We find that the deterioration in the 8th and 9th decimals is statistically significant, whereas the 7th decimal does not exhibit a significant change.

Moreover, our experiments indicate a negative correlation between the classifiers' yh parameter and the train-to-test precision deterioration, as shown in Table 8. The deterioration in yh = 1 and yh = 2 is statistically

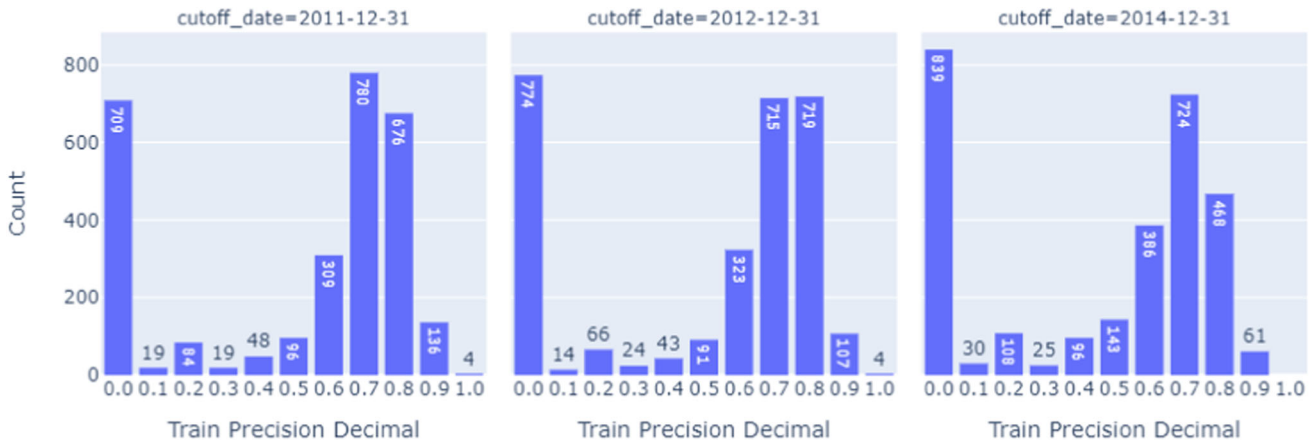

Fig. 3 Distribution of 2880 train precision results (per cut-off date)

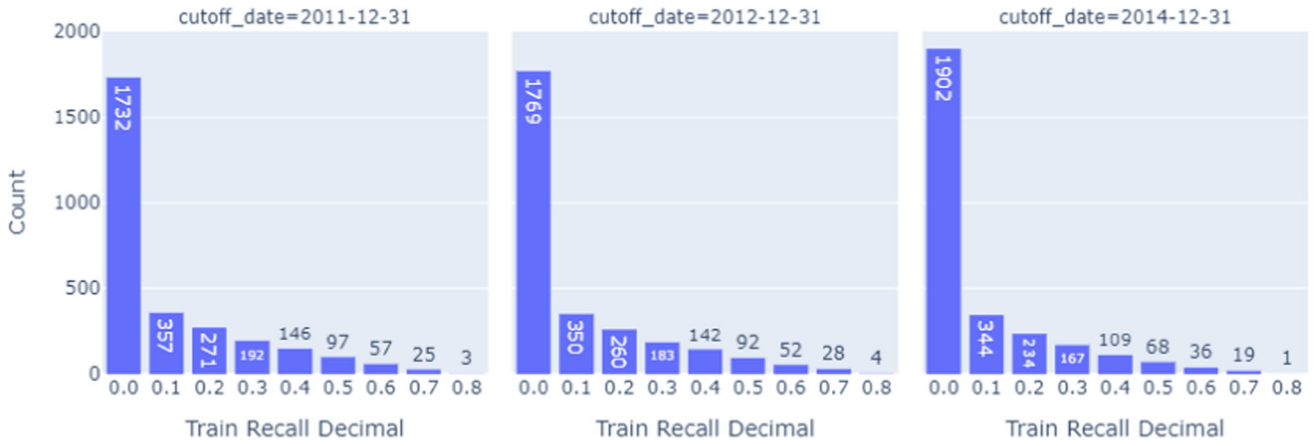

Fig. 4 Distribution of 2880 train recall results (per cut-off date)

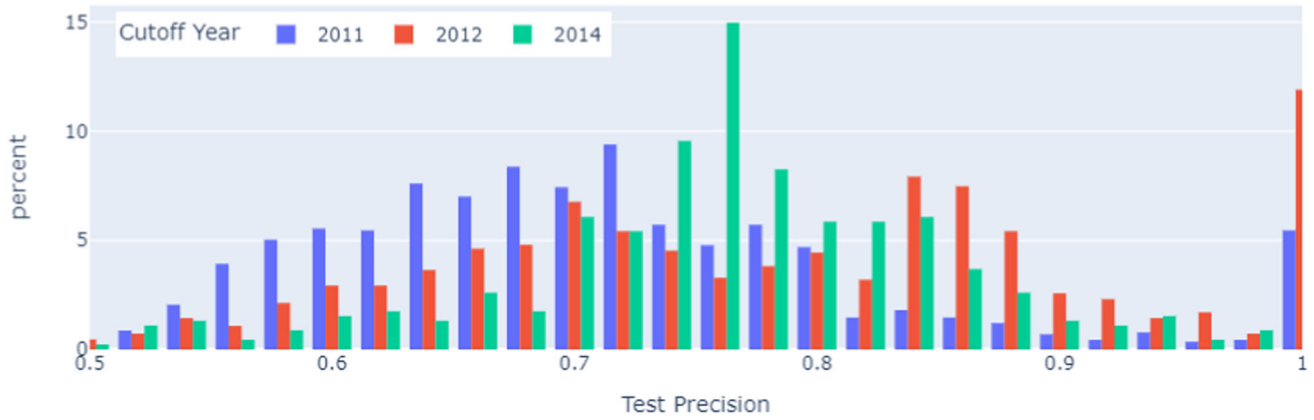

Fig. 5 Distribution of test precision (per cut-off year)

Table 5 Distribution of test recall results for classifiers with train precision result > 0.7 (per cutoff year)

Cutoff date	% of classifiers with train precision > 0.7	Test recall distribution		
		< 10%	10–30%	> 30%
31/12/2011	41%	21%	14%	6%
31/12/2012	39%	23%	12%	4%
31/12/2014	16%	11%	4%	1%

Table 6 Distribution of test recall results for classifiers with train & test precision result > 0.7 (per cut-off year)

Cutoff date	% of classifiers with test precision > 0.7	Test Recall range		
		0.0–0.1	0.1–0.3	0.3–1.0
31/12/2011	20%	10%	7%	3%
31/12/2012	27%	16%	8%	3%
31/12/2014	13%	9%	3%	1%

Table 7 Test versus train comparison per train precision decimal

Train precision	Test versus train mean % change	*p* value	Statistically significant
0.7–0.8	− 3.7%	0.4910	No
0.8–0.9	− 6.0%	0.0097	Yes
0.9–1.0	− 11.0%	0.0033	Yes

Table 8 test vs. train comparison per yh

yh	Mean test versus train % change	p value	Statistically significant
1	− 10.8%	∼ 0	Yes
2	− 6.8%	∼ 0	Yes
3	− 3.2%	1.0000	No

significant, while there is no significant deterioration observed for yh = 3.

Additionally, our experiments demonstrate that RF models consistently outperform GBM models by approximately 10% on average across all cutoff years (Fig. 6). This outperformance is statistically significant in all three cut-off years, as confirmed by the T-test of independent samples (p-values close to zero: $8.7 \cdot e^{-43}, 2.8 \cdot e^{-22}, 1.1 \cdot e^{-9}$). Furthermore, it appears that the models achieve better performance on average as the number of trees used increases (Fig. 7).

4.3 Test results: investment grade precision

Figure 8 quantifies the performance of the model in choosing investment grade stocks (recall these are stocks with minimum 15% AR within 3 years). The results are very encouraging, passing 80% in all cutoff years, nearing 90% in two of them.

Figure 9 shows that, as expected, the model's success rate at choosing investment grade stocks increases when the minimal training precision used is larger.

4.4 Computational costs

The scope of work described in the paper was implemented by a single 8-CPUs standard computer and a 40 Mbps download link to the Internet. The code's run time is as follows:

- Data Download (fundamentals & prices): < 1 h
- Pre-processing & features extraction: < 1 h
- Training of 144 classifiers (using 8 CPUs in parallel): < 4 h
- Predicting over test set: < 1 h

The total run time is < 7 h. Turning this model into a weekly-run production model should result in the same number of resources and run-time using the code as-is.

We also note that it might not be necessary to retrain the 144 models each week in a real production system, a lower frequency of retraining might be sufficient to maintain the models' performance. This should be checked during the development of such a production system and was added to the Future Work section of the manuscript. The reasons we suspect a lower frequency of retraining might be sufficient to maintain the models' performance are:

- The model (without any retraining) seems to output stable performance over long periods of test sets (recall that we measured the model's performance over test periods of 5–8 years using the same model trained on data from years before the test period)
- The amount of new training samples added each week equals the number of companies in the dataset. So, every week we add ∼ 2000 new data samples to a dataset of size 700–900 K (at the end of the research period). Such a low number of new weekly data samples might be less influential on the model's results versus the number of new data samples results from a lower frequency of retraining (e.g., monthly/quarterly retraining).

5 Discussion and future work

5.1 Features importance

5.1.1 Train process

Decision trees ensembles like RF and GBM quantify the importance of each feature to the performance of the

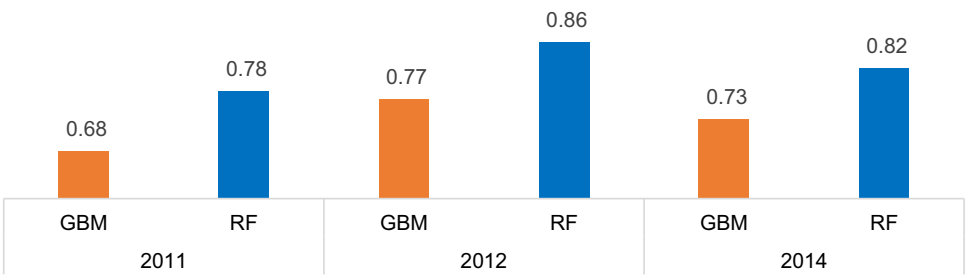

Fig. 6 RF vs. GBM mean test precision (per cutoff year)

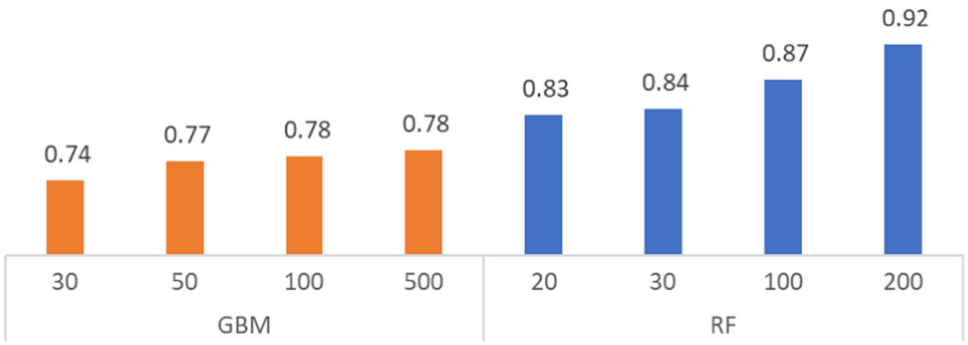

Fig. 7 Mean test precision per model type and number of trees used

Fig. 8 Mean investment grade precision per cutoff year

Fig. 9 Mean investment grade precision per cut-off year and minimal train precision (horizontal axis)

model. The calculation is based on the mean decrease in the impurity method used to select split points, like Gini (as in our case) or entropy. In simple words, the importance of a feature is a function of how much reduction in the impurity criterion was achieved as a result of all the splits the trees made based on that feature. While not a perfect method of features importance measurements, one could expect this method to point out the most important features in sufficient accuracy.

Analyzing the feature importance during the training phase of the classifiers, we observe that the value features significantly impact the classification decisions, contributing to an average feature importance of 80–85% for both RF and GBM models. Conversely, the influence of the quality features averages around 15–20% in the decision-making process, while the industry features appear to have limited importance. Figure 10 presents the features importance reported by one representative RF classifier.

5.1.2 Test process

We employed SHapley Additive exPlanations (SHAP) [43], who have gained significant attention in the field of machine learning for understanding feature importance. SHAP, a game-theoretic solution concept, estimates the contribution of features (known as SHAP values) by comparing model predictions with and without each feature in the dataset.

SHAP values provide a rigorous and mathematically grounded framework for attributing the contribution of each feature to the prediction of a machine learning model. The appeal of SHAP values lies in their ability to offer both global and local interpretability. Globally, they provide a comprehensive overview of the impact of each feature across the entire dataset, enabling practitioners to discern which features are consistently influential. Locally, SHAP values explain specific predictions on an individual level, shedding light on the reasoning behind each particular

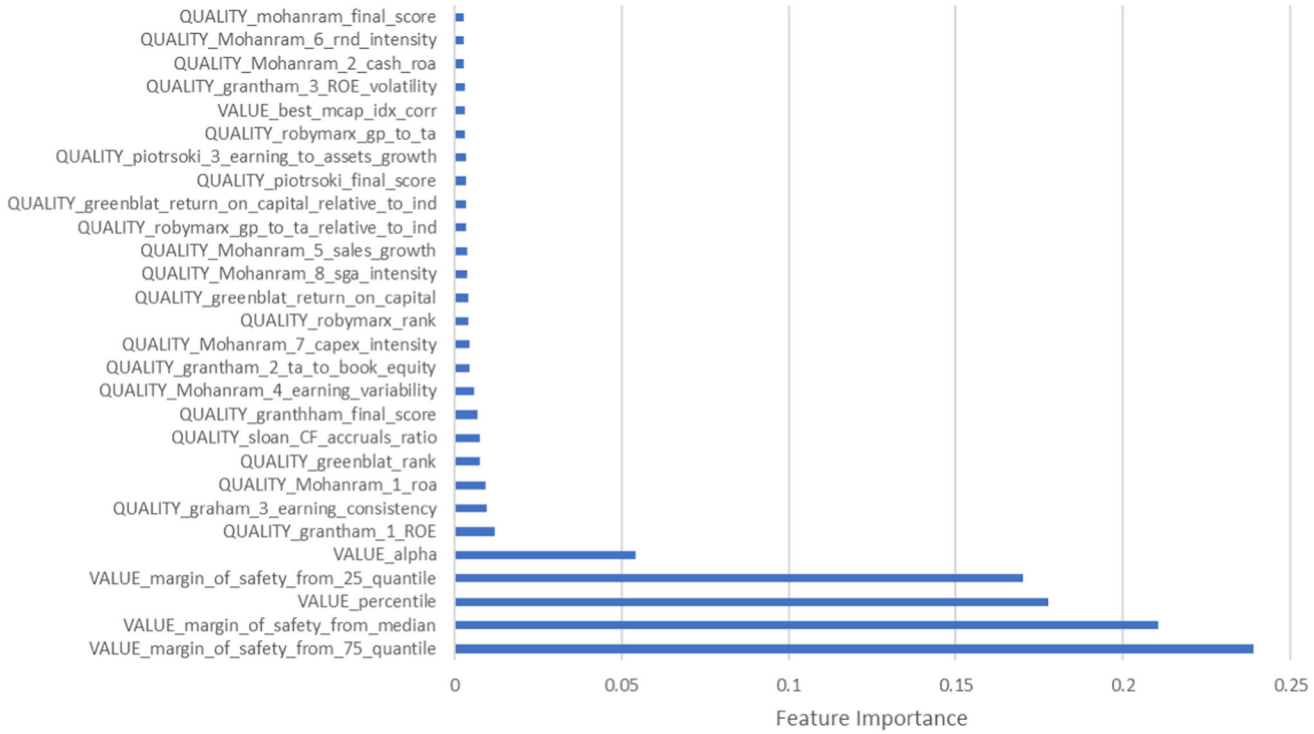

Fig. 10 Random Forrest features importance

outcome. This dual perspective empowers researchers and practitioners to not only grasp the overall behavior of their models but also to delve into the intricacies of individual predictions.

Furthermore, the application of SHAP values is agnostic to the underlying machine learning model. Whether dealing with linear regression, complex ensemble methods, or sophisticated deep learning architectures, SHAP values can be seamlessly employed, ensuring a consistent approach to feature importance analysis across diverse models. Additionally, SHAP values inherently account for feature interactions, a critical aspect of many real-world problems. By considering all possible combinations of features, SHAP values capture subtle dependencies and interactions, providing a nuanced understanding of how features influence predictions in concert.

In Fig. 11, we present the top 20 features with the highest SHAP values obtained from a single RF test classifier. These features are ranked in descending order of importance. The "SHAP Value (Impact on model output)" horizontal axis signifies whether the feature has a positive or negative effect on the prediction. The color scheme, with red representing high values and blue representing low values, indicates the corresponding variable values for each observation.

The SHAP results clearly find the margin-of-safety value features to be the most influential ones, with higher *mos* (red points) associated with an increased likelihood of being deemed a good value investment (positive SHAP values), whereas a lower *mos* (blue points) is associated with a reduced likelihood of being considered a good value investment (negative SHAP values). This aligns with our initial expectations, considering that *mos* is a part of the class definition, as well as prove consistent with the train-process feature importance presented at Sect. 5.1.1

The remaining impact on the model's predictions stems from the Quality features, which fine-tune the algorithm's predictions. Interestingly, the Industry features exhibit minimal impact on the model's predictions. This again is consistent with the train process features importance we presented.

5.2 Model's sensitivity to class_type input parameters (mos & yh)

As previously mentioned, the margin of safety serves as a safeguard for value investors against their own valuation judgment errors. A larger margin of safety implies that valuation mistakes must be substantial to significantly impact future returns. Figure 12 demonstrates that our model adheres to this trend. Specifically, in two out of the three cutoff years (the ones with longer training periods), increasing the margin of safety from 0.3 to 0.4+ leads to a significant rise in investment grade precision, elevating it from approximately 80% to above 90%.

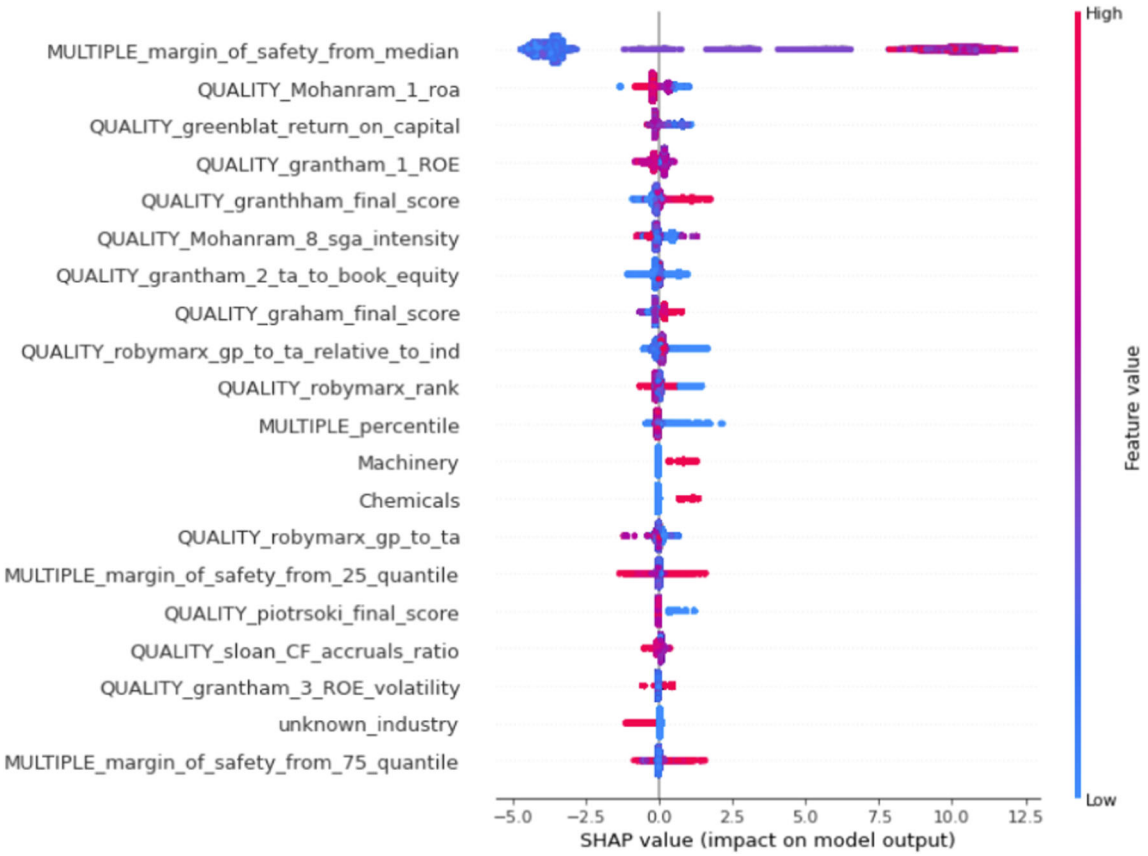

Fig. 11 The top 20 SHAP values of a RF test classifier

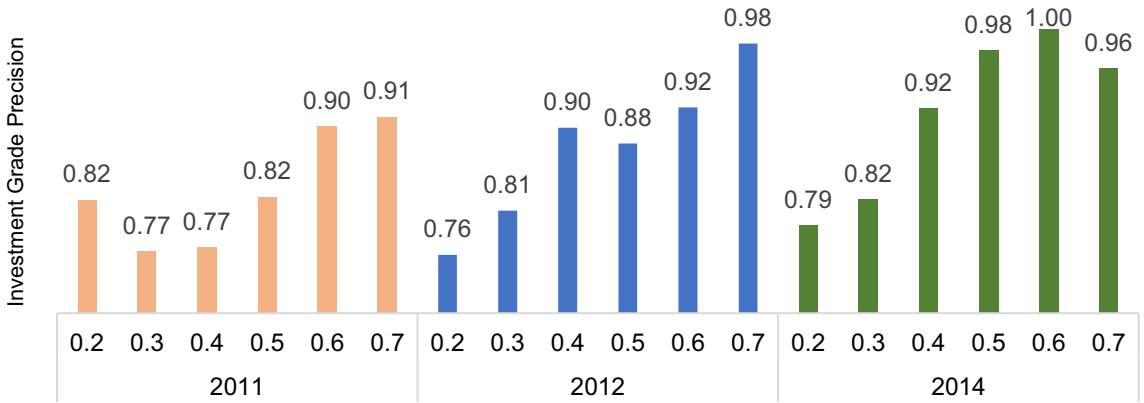

Fig. 12 Model's mean investment grade precision per cutoff year and class's mos (horizontal axis)

Moreover, the likelihood of a successful value investment increases as we consider a larger value for the years ahead (yh) parameter, as this allows more time for the stock to converge with its intrinsic value. Figure 13 provides evidence supporting this notion. In two of the three cut-off years (again, the ones with longer training periods), transitioning from yh = 1 to yh = 2 years results in a substantial increase in the success ratio of investment grades, escalating it from around 80% to above 90%.

Additionally, we anticipate a reduction in the number of stocks recommended by our model as we impose stricter conditions, such as higher training precision, larger margin of safety, and longer years ahead. Figure 14 corroborates this expectation, focusing on cutoff year = 2012 (similar trends can be observed in the other two cutoff years). (Please note the logarithmic scale of the vertical axis.) A clear downward trend is visible as the margin of safety increases, holding true for various yh values and minimal

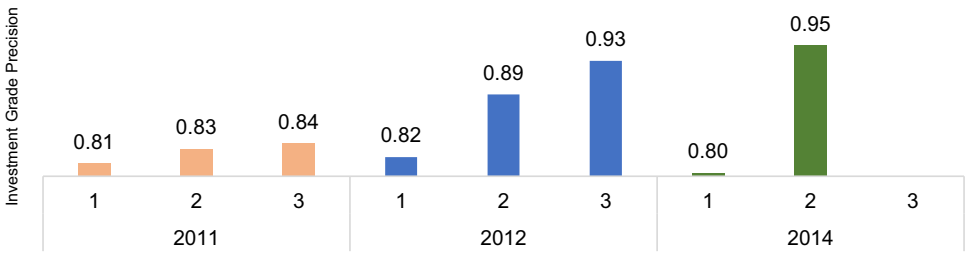

Fig. 13 Model's investment grade precision per cutoff year and class's yh (horizontal axis)

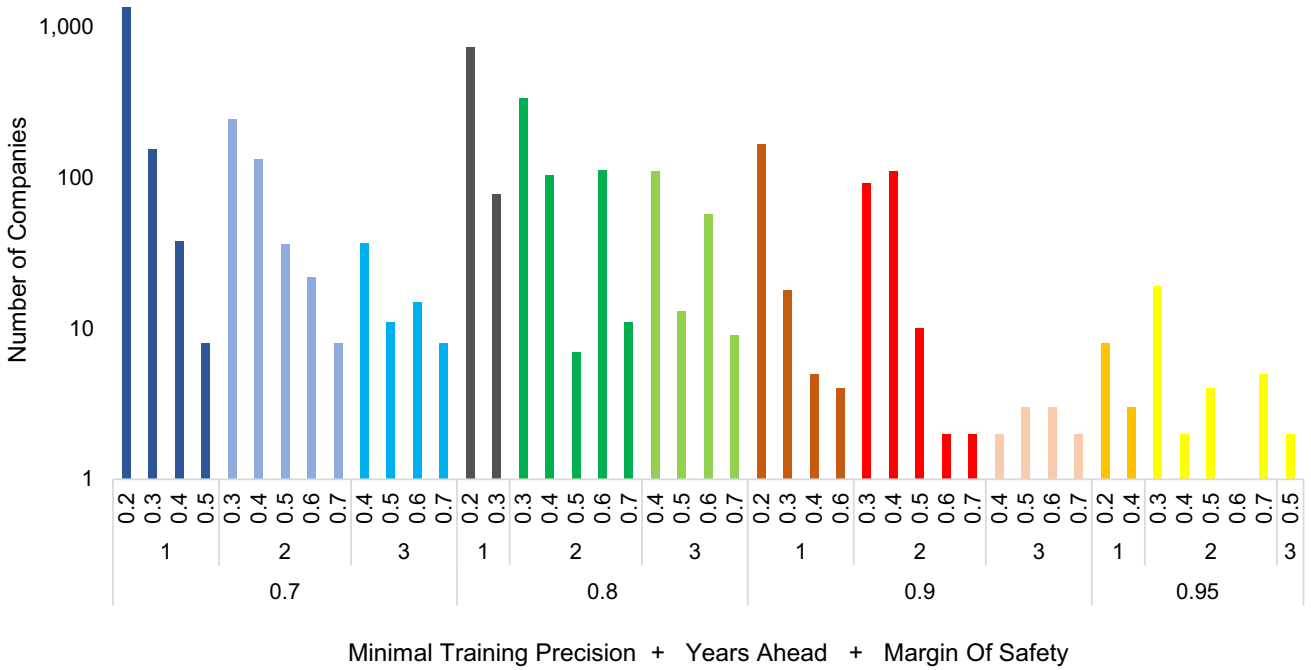

Fig. 14 Number stocks recommended by the model distribution (cutoff year = 2012)

training precision levels (evident from the declining trend in each distinct color on the chart). Furthermore, as yh increases, a similar downtrend can be observed (exemplified by the three blueish color groups, representing the three yh values below the 0.7 minimal training precision threshold). Lastly, a noticeable decline is evident as the minimal training precision increases (highlighted by the difference in scale between the greenish column group with a minimal training precision of 0.8, the reddish group with a minimal training precision of 0.9, and the yellowish group with a minimal training precision of 0.95).

5.3 Chosen stocks distribution by industry and size

Value investing has historically demonstrated stronger performance over companies with lower market value, which are more susceptible to significant market mispricing, particularly in stable and predictable industries. To visualize the distribution of stocks across the market value size dimension, Fig. 15 showcases the entire dataset alongside the stocks selected by the model in cutoff year = 2012. The percentage of small market value stocks (including 'small,' 'micro,' and 'nano' size categories) in the entire dataset is 54%, whereas the model chose 68% such smaller stocks. This disparity in the proportion of smaller stocks was confirmed to be statistically significant (The chi-square good-of-fitness test was used to assess the null hypothesis that the stocks chosen by the model have the same size distribution as the whole dataset, versus the alternative hypothesis that they do not. The values $p < 0.05$ were considered statistically significant. The test resulted with chi-square statistic of 359.49, with p value of $1.57 \bullet e^{-75}$).

We conducted a similar chi-square test to examine whether the stocks chosen by the model differ in terms of their industry distribution compared to the entire dataset. The test yielded a statistically significant result (p value of 0.0286). However, from a practical investment perspective, this discrepancy is less useful, with a maximum difference of 2%.

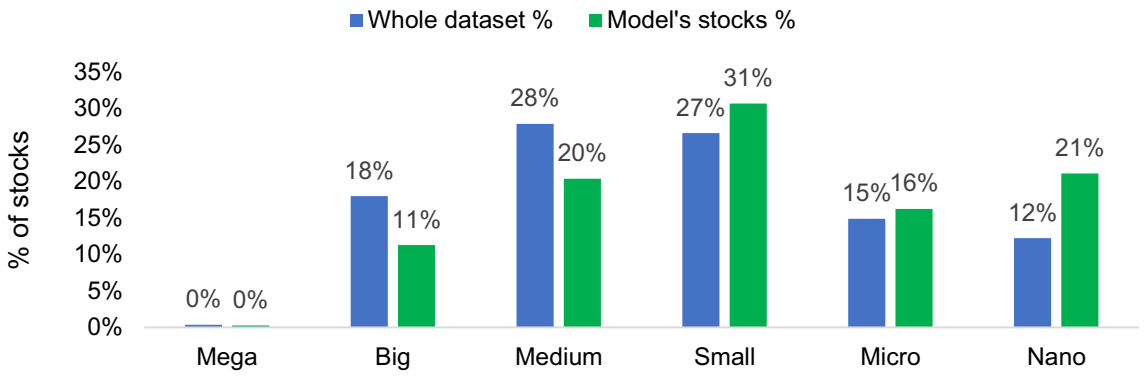

Fig. 15 Distribution of stocks per size category—whole dataset and model's output (cutoff year = 2012)

5.4 Do we have less stocks chosen when markets are up?

We explore whether there are fewer stocks chosen by the model when the markets are experiencing an upward trend. Since the definition of a value stock is partially influenced by its current market price, it follows that the number of value stocks diminishes as equity market prices increase, and vice-versa. For example, many value investors have emphasized that the stock market in early 2009, following the September 2008 financial crash, presented unprecedented buying opportunities.

Our longest test period corresponds to cutoff year = 2011, encompassing a full 7-year testing period of 2013 to 2019 (inclusive). This period, in general, exhibited a bullish trend for the Russell-3000 index, with only two short sub-periods of significant market declines in Q1-2016 and Q4-2018.

As depicted in Fig. 16, the number of recommended stocks witnessed a remarkable surge during the market decline in Q1-2016, with the monthly stocks in January and February 2016 being 6–8 times higher than the average monthly stocks chosen in the preceding six months and the subsequent four months. A similar increase in the number of stocks selected by the model is evident during the Q4-2018 period, where the number of stocks is approximately double the average quarterly stocks chosen in Q1–Q3 2018.

5.5 Portfolio example

For our stock picking strategy to be practically valuable, it must be translated into a managed portfolio that involves a multitude of decisions, such as determining the initial capital to invest, the number of stocks to include, the allocation of cash to each stock, considerations for diversification, stock replacement and selling policies, and more. While portfolio construction and management is not the main focus of our research, we do provide an example of a simple portfolio based on our model's stock picks.

It is evident that the results we present can be further improved by optimizing the portfolio management policy we used. This policy is based on the following straightforward rules:

- The model's input parameters are as follows:
  - Minimal training precision = 0.8
  - Minimal class's AR = 0.1
  - Cutoff year = 2012
  - Max years ahead used in training = 1

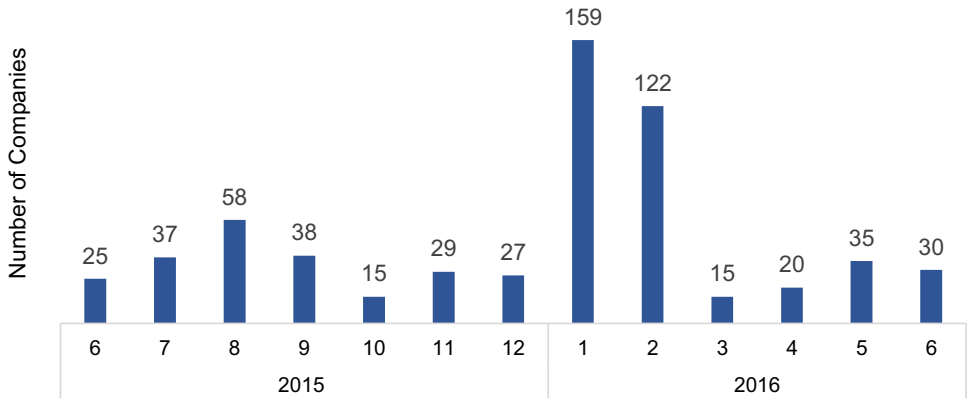

Fig. 16 Number of monthly stocks chosen by the model during June 2015–June 2016 (cutoff year = 2012)

Fig. 17 1M$ investment in our portfolio (blue) versus Russell-3000 (red), 2014–2019

- Based on the aforementioned choices, the first possible buy date is 1/1/2014, as the training period ends on 31/12/2012 and we utilize a maximum of 1 year of data ahead for class calculation, rendering the 2013 data unusable.
- We terminate our portfolio by selling all stocks at market price on 31/12/2019, resulting in a portfolio lifespan of 6 years.
- Our initial capital is $1 million, and we allocate $100,000 per company. Thus, we aim for a concentrated portfolio consisting of 10 stocks. Although this number may fluctuate over time based on portfolio performance, its magnitude will remain relatively stable.

Regarding the buying policy:

- We receive weekly inputs from the model.
- For each stock, we consider only the first time it is recommended by the model.
- On a weekly basis, we sort all recommended stocks in descending order of expected AR, as determined by the class of the classifier that raised the recommendation. If multiple classifiers recommend the same stock in the same week, we select the class with the highest AR.
- After sorting, we purchase the stocks in the specified order, one by one, as long as our cash reserves allow. If there is insufficient cash, we dismiss the stock, and it will not be included in the portfolio at all.

Regarding the selling policy:

- The maximum holding period for a stock is defined as 2 years.
- On a weekly basis, we monitor the price of each purchased stock and sell it based on one of the following two scenarios:
  - If a stock's return (compared to the purchase price) exceeds 15%, we establish an "upward threshold" initially set at 15%. If the return grows to 20%, 25%, 30%, and so on, we increase the threshold accordingly. We sell the stock if its current weekly price falls 5% below the latest upward threshold. For example, if the latest upward threshold is 25%, we sell the stock if the price reflects a return lower than 20%.
  - If a stock's return does not reach 15% within 2 years, we sell the stock.

Based on these portfolio management decisions, the model recommended a total of 326 companies, of which only 122 were included in the final portfolio spanning 6 years due to insufficient cash to purchase the remaining 204 stocks. As a result, we have a concentrated portfolio with an average of approximately 20 concurrently held stocks. Approximately 50% of the stocks are sold within 6 months, 25% within a year, and the remaining 25% within 1–2 years. Figure 17 illustrates the performance of our portfolio compared to the Russell-3000 index. Our portfolio achieved an annualized return of 15%, surpassing the index's 9%. However, as expected from a concentrated portfolio like ours, the portfolio's volatility is more than double that of the index, resulting in a lower Sharpe ratio of 0.52 compared to the index's 0.64.

5.6 Conclusions and future work

5.6.1 Conclusions

Our research has demonstrated the effectiveness of utilizing traditional value and quality investing features in conjunction with weekly evaluations of thousands of companies, resulting in a high-performing machine learning model for stock picking. Surpassing our initial expectations, this model has achieved an impressive picking rate of over 80% for investment grade stocks while maintaining high enough recall performance to ensure that the number of stocks picked is sufficient.

The sample portfolio we composed, while using straightforward portfolio management policy, capture the essence of real0world portfolio management considerations, such as when to buy and sell a stock, and allocation of funds among stocks. In this aspect, the superior performance of the model vs. the Russel-3000 over the test period encourages real-world usage of our model.

5.6.2 Future work

It is evident to us that there are several avenues for further improvement in this model:

- Enhancing the current model architecture by optimizing the hyperparameters of the existing random forest (RF) and gradient boosting machine (GBM) algorithms, as well as incorporating additional classification algorithms such as neural networks (NN) or support vector machines (SVM).
- Expanding the dataset by acquiring data from non-US stock exchanges and obtaining a more extensive historical dataset for US stocks. This broader data coverage can enhance the model's ability to capture global market dynamics and improve its overall performance.
- Addressing the issue of missing data in our current dataset, which has led to the exclusion of numerous companies from consideration. Utilizing a more comprehensive and well-maintained dataset, such as Compustat, could potentially alleviate this problem and mitigate the survivorship bias present in our final universe of stocks.
- Incorporating additional quality features into the model to enhance its ability to differentiate between value investing stocks and those that are simply priced cheaply.
- Optimizing the portfolio construction process, as mentioned earlier, to maximize the performance and risk-reward characteristics of the final portfolio. This involves considering factors such as diversification, more advanced portfolio risk management methods, differentiated allocation of initial funds to different stocks as function of the model's prediction strength, and the inclusion of dynamic portfolio rebalancing strategies.
- As we noted in Sect. 4.4, it might not be necessary to retrain the 144 models each week in a real production system, a lower frequency of retraining might be sufficient to maintain the models' performance. This should be checked during the development of such a production system.

Performing a features selection process, to focus on a subset of current model's features, to reduce train time and to potentially improve the model's prediction accuracy. We performed an initial analysis using only the top 7 features described in Sect. 5.1 and found that the training process could be accelerated by magnitude of dozens of percentages without any notable deterioration in the models' prediction accuracy. A proper features process selection should be added in the future to our train process, using a separate validation set.

Appendix A: Financial terms glossary

The following presents the definition of the financial and accounting terms used in our research [44]:

Term (acronym)	Definition
Annualized Return (AR)	Annualized return is the average rate of return an investment generates per year over a specified period of time. It is calculated to provide investors with a normalized performance metric, especially when comparing investments with different time frames. The annualized return takes into account the effects of compounding, giving investors a more accurate measure of an investment's profitability over time
Asset Turnover	Asset turnover is a financial ratio that measures a company's efficiency in using its assets to generate sales revenue. It is calculated by dividing the total sales revenue by the average total assets. A higher asset turnover ratio indicates that a company is effectively utilizing its assets to generate sales, which is a positive sign for investors
Balance Sheet	A balance sheet is a financial statement that provides a snapshot of a company's financial position at a specific point in time. It shows the company's assets, liabilities, and shareholders' equity. The balance sheet provides investors and analysts with insights into a company's financial health, including its liquidity, solvency, and overall wealth

Term (acronym)	Definition
Book Value (BV)	Book value, also known as shareholders' equity or net asset value, is the total value of a company's assets that shareholders would theoretically receive if a company were liquidated. It is calculated by subtracting a company's total liabilities from its total assets. Book value per share is a common metric used by investors to assess a stock's valuation
Cash Flow from Operations (CFO)	Cash flow from operations represents the cash generated or used by a company's normal business operations. It provides insights into a company's ability to generate cash from its core activities. Positive cash flow from operations indicates that a company can sustain and grow its business, while negative cash flow may raise concerns about its financial health
Cashflow Statement	A cash flow statement is a financial statement that provides information about a company's cash inflows and outflows over a specific period. It categorizes cash flows into operating, investing, and financing activities. The cash flow statement helps investors assess a company's ability to generate cash and manage its liquidity, providing valuable insights into its financial stability
Current Ratio	The current ratio is a liquidity ratio that measures a company's ability to cover its short-term obligations with its short-term assets. It is calculated by dividing current assets by current liabilities. A current ratio above 1 indicates that a company has more current assets than current liabilities, suggesting strong liquidity and the ability to meet short-term obligations
Discounted Cash Flows (DCF)	Discounted cash flows (DCF) is a valuation method used to estimate the value of an investment based on its future cash flows. DCF analysis calculates the present value of projected cash flows by discounting them back to their current value using a discount rate. DCF is widely used in finance for valuing investments, businesses, and projects, providing a net present value (NPV) perspective
Dividends	Dividends are payments made by a corporation to its shareholders, usually in the form of cash or additional shares of stock. Dividends are typically paid out of a company's profits as a way to distribute earnings to shareholders. Dividend payments are an important source of income for investors, especially those seeking stable and consistent returns from their investments
Earnings Per Share (EPS)	Earnings per share (EPS) is a financial metric that represents the portion of a company's profit allocated to each outstanding share of common stock. It is calculated by dividing net income by the number of outstanding shares. EPS is a key indicator of a company's profitability and is often used by investors to assess a company's financial performance on a per-share basis
Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA)	EBITDA is a financial metric that represents a company's earnings before deducting interest, taxes, depreciation, and amortization expenses. EBITDA provides insights into a company's operational profitability and cash flow from its core business activities, excluding the impact of non-operating items and financial decisions. Investors use EBITDA to assess a company's operating performance and financial health
Enterprise Value (EV)	Enterprise value (EV) is a measure of a company's total value, representing its market capitalization plus its total debt and minority interest, minus its cash and cash equivalents. EV provides a more comprehensive view of a company's value, considering both equity and debt. It is often used in financial analysis to compare the value of different companies on an apples-to-apples basis

Term (acronym)	Definition
Enterprise Value/EBITDA ratio	The Enterprise Value/EBITDA ratio is a financial metric that compares a company's enterprise value to its earnings before interest, taxes, depreciation, and amortization (EBITDA). It is used to assess a company's valuation relative to its EBITDA, providing insights into its financial performance and potential profitability. A lower ratio may indicate an undervalued company
Gross Margin (GM)	Gross margin is a profitability ratio that measures the percentage of revenue a company retains after deducting the cost of goods sold (COGS). It is calculated by dividing gross profit by total revenue and multiplying by 100 to get a percentage. Gross margin indicates how efficiently a company produces and sells its products, providing insights into its pricing strategy and cost structure
Income Statement	An income statement, also known as a profit and loss statement, is a financial statement that shows a company's revenues, expenses, gains, and losses over a specific period. It provides insights into a company's profitability by detailing its revenues and the costs associated with generating those revenues. The income statement is a crucial tool for financial analysis
Investment Intensity	Investment intensity is a financial metric that measures a company's capital investment in relation to its sales revenue. It is calculated by dividing capital expenditures by total sales revenue. Investment intensity provides insights into a company's investment strategy and capital allocation, helping investors assess its commitment to future growth and expansion
Leverage Ratio	The leverage ratio measures a company's level of debt relative to its equity and other financial metrics. It provides insights into the company's financial risk and ability to meet its debt obligations. The leverage ratio is calculated in various ways, such as the debt-to-equity ratio, providing investors with information about the company's capital structure and financial stability
Market Value (MV)	Market value, also known as market capitalization, is the total value of a company's outstanding shares of stock. It is calculated by multiplying the current stock price by the total number of outstanding shares. Market value represents the market's perception of a company's worth and is a key indicator used by investors to assess a company's size and relative importance in the market
Multiples	Multiples refer to various financial ratios, such as price-to-earnings ratio, price-to-book ratio, and enterprise value-to-EBITDA ratio, used to compare a company's valuation to its earnings, book value, or EBITDA. Multiples provide insights into how the market values a company relative to its financial performance and industry peers, aiding investors in making investment decisions
Net Income	Net income, also referred to as net profit or earnings, is the total amount of money a company has earned or lost during a specific period after all its expenses, taxes, and other financial activities have been deducted from its total revenue. Net income provides a clear indication of a company's profitability and is a key figure used by investors and analysts to assess a company's financial performance and potential for growth
Price/Book ratio (P/B, also MV/BV)	The price-to-book (P/B) ratio is a financial metric that compares a company's market price per share to its book value per share. It is calculated by dividing the current market price per share by the book value per share. P/B ratio provides insights into how the market values a company's assets in relation to its market price, indicating whether the stock is undervalued or overvalued

Term (acronym)	Definition
Price/Earnings ratio (PE)	The price-to-earnings (P/E) ratio is a valuation ratio that compares a company's market price per share to its earnings per share (EPS). It is calculated by dividing the current market price per share by the earnings per share. P/E ratio helps investors assess a stock's valuation and future earnings potential. A higher P/E ratio may indicate a higher expectation of future earnings growth
Return on Assets (ROA)	Return on assets (ROA) is a financial ratio that measures a company's ability to generate profit from its total assets. It is calculated by dividing net income by average total assets. ROA indicates how efficiently a company utilizes its assets to generate earnings. A higher ROA suggests that a company is effective in converting its investments in assets into profits, which is favorable for investors
Return on Equity (ROE)	Return on equity (ROE) is a financial ratio that measures a company's profitability and efficiency in generating profits from its shareholders' equity. It is calculated by dividing net income by shareholders' equity. ROE indicates how well a company utilizes shareholders' funds to generate profits. A higher ROE suggests that a company is using its equity effectively to generate profits, which is a positive sign for investors
Return On Invested Capital (ROIC)	Return on invested capital (ROIC) is a financial metric that measures a company's ability to generate profit from its invested capital, including both equity and debt. It is calculated by dividing net operating profit after taxes by the average invested capital. ROIC provides insights into a company's efficiency in utilizing its capital to generate returns for shareholders and debt holders
Sloan Accruals	Sloan accruals refer to a method used to identify companies that may be manipulating their earnings. It involves analyzing the relationship between a company's reported earnings and its cash flows. If a company's reported earnings are significantly higher than its cash flows, it may suggest aggressive accounting practices, potentially raising concerns among investors and analysts about the company's financial transparency and integrity
Stock Buybacks	Stock buybacks, also known as share repurchases, occur when a company buys back its own shares from the open market. Companies may engage in stock buybacks to reduce the number of outstanding shares, increase earnings per share (EPS), or return excess cash to shareholders. Stock buybacks can impact a company's stock price and financial ratios, influencing investor perceptions and investment decisions
Total Assets/Long-Term Debt	Total assets to long-term debt ratio is a financial metric that compares a company's total assets to its long-term debt. It provides insights into a company's ability to cover its long-term debt obligations with its total assets. A higher ratio indicates a stronger financial position, as the company has more assets to cover its long-term debt, reducing the risk of financial distress and default

Appendix B: Hyperparameters used in the prleiminary analysis

The following table summarizes the different hyperparameters tuned with multiple values during the preliminary analysis described in Sect. 3.6.1:

Algorithm	Hyperparameters
Random Forrest	• Number of trees • Maximum depth of a tree • class_weight (None/Balanced in python scikit-learn package)
Gradient Boosting Machine	• Number of trees • Maximum depth of a tree • class_pos_weight (in python Catboost package) • Tree grow policy (Symmetric/Lossguide in python Catboost package)
Logistic Regression	• Regularization (l1/None in python scikit-learn package)
Deep Learning	• Number of hidden layers • Number of Neurons per hidden layer

Funding Open access funding provided by Ben-Gurion University. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The raw datasets are purchased from a 3rd party vendor (QuickFS and FMP) and are hence not publicly available. The dataset used for the modeling itself—including the value and quality features as well as the classes labels—is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interest to disclose.

References

1. IFRS (2022) IFRS Issued Standards. [Online]. Available: https://www.ifrs.org/issued-standards/list-of-standards/
2. FASB (2022) FASB Home Page. [Online]. Available: https://www.fasb.org/home
3. Graham B, Dodd D (1934) Security Analysis, New York. Whittlesey House, New York
4. Graham B (1973) The intelligent investor, 4th Revised. Harpers & Row, New York
5. Klarman S (1991) Margin of safety: risk-averse value investing strategies for the thoughtful investor. Harper-Collins, New York
6. Greenblatt J (2010) The little book that still beats the market. Wiley, Hoboken
7. Van Den Berg A (2020) Ivey Value Investing Classes Guest Speaker: Arnold Van Den Berg, Ivey Business School, 2020. [Online]. Available: https://www.youtube.com/watch?v=Uu4EQTPaQdE&list=PL06tLdGWaCMpZsPHL9PaeJqu-s5cKvikW&index=6&t=1279s
8. Bains H (2021) Ivey Value Investing Classes Guest Speaker: Hardev Bains, Ivey Business School, 2021. [Online]. Available: https://www.youtube.com/watch?v=FFT3aJoWAd4&list=PL06tLdGWaCMpZsPHL9PaeJqu-s5cKvikW&index=1
9. Melvin T (2013) 8 Rules for Picking Perfect Value Stocks, Marketfy, 2013. [Online]. Available: https://www.youtube.com/watch?v=ayG6h9d2Cb8&list=PL06tLdGWaCMpZsPHL9PaeJqu-s5cKvikW&index=2
10. Lynch P (1989) One up on Wall Street, New York. Simon and Schuster, New York
11. Greenblatt J (2019) 2 Secrets to Beating the Market, 6 April 2019. [Online]. Available: https://www.youtube.com/watch?v=79Y-cc2nVdc&list=PL06tLdGWaCMpZsPHL9PaeJqu-s5cKvikW&index=11
12. Klarman S (2009) 2009 Ivey Value Investing Classes Guest Speaker: Seth A. Klarman, Ivey Business School. [Online]. Available: https://www.youtube.com/watch?v=8Shv9k6UU00
13. Greenwald BC (2004) Value investing: from graham to buffett and beyond. Wiley, Hoboken
14. Phillips S (2012) Buying at the point of maximum pessimism. FT Press, Upper Saddle River
15. Koller T, Goedhart M, Wessels D (2004) Valuation: measuring and managing the value of companies, 4th edn. Wiley, Hoboken
16. Damodaran A (2012) Investment valuation: tools and techniques for determining the value of any asset, 3rd edn. Wiley, Hoboken
17. MSCI (2013) Foundations of Factor Investing. MSCI
18. Basu S (1977) Investment performance of common stocks in relation to their price-earnings ratios: a test of the efficient market hypothesis. J Finance 32(3):663–682
19. Banz R (1981) The relationship between return and market value of common stocks. J Financ Econ 9(1):3–18
20. Fama E, French K (1993) Common risk factors in the returns. Journal of Finance 33(1):3–56
21. Fama E, French K (2015) A five-factor asset pricing model. J Financ Econ 116(1):1–22
22. Harvey C, Liu Y, Zhu H (2015) … and the cross-section of expected returns. Rev Financ Stud 29(1):5–68
23. Novy-Marx R (2016) Quality investing. University of Rochester. Working paper
24. Sharp W (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Finance 19(3):425–442
25. Fundsmith (2021) Annual Letter to Shareholders, Fundsmith LLP
26. O'Shaughnessy P (2016) Combining the best stock selection factors. In: QuantCon, New York City

27. Olson D, Mossman C (2003) Neural network forecasts of Canadian stock returns using accounting ratio. Int J Forecast 19(3):453–365
28. Cao Q, Leggio K, Schniederjans M (2005) A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. Comput Oper Res 32(10):2499–2512
29. Kryzanowski L, Galler M, Wright D (1993) Using artificial neural networks to pick stocks. Financ Anal J 49(4):21–27
30. Abe M, Nakayama H (2018) Deep learning for forecasting stock returns in the cross-section. In: Pacific-Asia conference on knowledge discovery and data mining
31. Alberg J, Lipton Z (2017) Improving factor-based quantitative investing by forecasting company fundamentals. arXiv preprint arXiv:1711.04837
32. Yang H, Liu X-Y, Wu Q (2018) A practical machine learning approach for dynamic stock recommendation. In: 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference
33. Quick FS, [Online]. Available: www.quickfs.com
34. Financial Modeling Prep, [Online]. Available: https://site.financialmodelingprep.com/
35. S&P (2018) GICS-Global Industry Classification Standard. [Online]. Available: https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook_2018_v3_letter_digitalspreads.pdf
36. Piotroski J (2000) Value investing: the use of historical financial statement information to separate winners from losers. J Account Res 38:1–41
37. Novy-Marx R (2013) The other side of value: the gross profitability premium. J Financ Econ 108(1):1–28
38. Mohanram P (2005) Separating winners from losers among lowbook-to-market stocks using financial statement analysis. Rev Acc Stud 10:133–170
39. GMO, "The Case for Quality - the Danager of Junk," GMO white paper, 2004.
40. Sloan R (1996) Do stock prices fully reflect information in accruals and cash flows about future earnings? Account Rev 71(3):289–315
41. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
42. Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232
43. Lundberg S, Su-In L (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774
44. Investopedia, [Online]. Available: https://www.investopedia.com/terms/m/marketcapitalization.asp