Beyond Time: Accurately Estimating the Fair Value of Stocks with Machine Learning and Fundamental Data

Daniel Netzl

Abstract

This paper presents a novel machine learning methodology to assess the intrinsic value of firms, based on both qualitative and quantitative data, and deliberately excluding time series analysis. The employed models demonstrate proficiency in identifying undervalued stocks, highlighting the importance of effective feature engineering and data analysis in finance. This method represents a potential departure from traditional stock valuation techniques, suggesting a new direction in investment strategies. While results require careful interpretation, they indicate a potential shift in investment paradigms.

1 Introduction

This paper addresses the tendency of investors, particularly younger ones, to prioritize quick gains over methodical decision-making. With the recent surge in financial markets attracting many inexperienced investors, often influenced by online speculation, the need for a data-driven approach to asset management is clear. This trend has led to risky investment behaviors, exacerbated by easy access to online trading platforms. This paper advocates for a shift from speculative to informed investing by introducing a machine learning algorithm that utilizes both qualitative and quantitative data to estimate a company's intrinsic value. The study examines the model's ability to accurately assess firm value and its effectiveness in outperforming the market.

2 Related Work

The field of study predominantly focuses on high-frequency trading and short-term price fluctuations, with less emphasis on long-term, stable solutions. The integration of qualitative metrics in machine learning for stock evaluation is limited, indicating a gap in current research.

Recent studies have significantly contributed to the field of machine learning in stock market analysis. For instance, Chen et al. (2019) successfully applied deep neural networks to asset pricing, achieving superior performance over traditional methods. Milosevic (2016) focused on long-term stock price forecasting using current financial data, achieving a notable 76.5% F-score, which signifies a strong balance between precision and recall in the model's predictive accuracy. And, lastly, Amel-Zadeh et al. (2020) provided a foundational understanding of how neural networks and random forest regression could be harnessed to predict stock returns. Their investigation, which highlighted the strengths and limitations of these models in financial contexts, offered guidance on model selection and feature importance, directly informing the methodological choices in the current study.

Distinct from these works, the current study concentrates on long-term value investing, seeking to identify undervalued stocks. By focusing on a broad set of parameters and excluding the time-series dimension, this approach minimizes the noise and potential overfitting associated with short-term fluctuations, allowing for a clearer assessment of a stock's sustainable value.

3 Methodology

3.1 Data Collection and Sources

This study harnesses a rich dataset from 396 dividend-distributing companies, showcasing a blend of both qualitative and quantitative metrics. The data was meticulously gathered from Financial Modeling Prep (FMP), an API offering a comprehensive range of financial data, alongside other sources such as social media sentiment, insider trading activities, and Environment, Social, and Governance (ESG) scores. This approach enables a holistic view of each company beyond traditional financial statements. The incorporation of 23 broader market parameters like GDP and unemployment rates provides additional context to the stock evaluations.

3.2 Data Preparation and Feature Engineering

The initial dataset encompassed 337 predictors across 4,163 entries, after applying stringent data completion requirements. Entries with substantial missing data were excluded to maintain the integrity of the analysis.

3.2.1 Feature Selection

Feature selection was conducted through a multi-step process to mitigate multicollinearity and enhance model performance. Variables exhibiting a Pearson correlation coefficient above 0.95 were identified as highly correlated and thus candidates for removal to reduce redundancy. To further address multicollinearity, the Variance Inflation Factor (VIF) was employed to quantify the increase in model variance for one unit increase in the predictor's variance, with a threshold set to 5 to identify and exclude highly collinear variables. As the final step in feature selection, Principle Component Analysis (PCA) was utilized to transform the data into a set of linearly uncorrelated variables, known as principal components. This method significantly reduced dimensionality while retaining the variance in the dataset, facilitating more efficient and effective model training.

3.2.2 Stock Evaluation Formulae

To determine the intrinsic value of stocks, this study employs three renowned valuation techniques. These methods, grounded in financial analysis, assess the worth of a company from different perspectives, using both its current financial health and potential future earnings. For detailed formulae and computation methodologies, refer to Appendix Section A.1.

The Discounted Cash Flow (DCF) method values a company by projecting its future cash flows and discounting them to present value, reflecting the premise that the value of a company lies in its ability to generate cash in the future. This method emphasizes the time value of money and the importance of cash flows over mere earnings, making it particularly suited for evaluating companies with significant future growth potential.

Inspired by Benjamin Graham, known as the father of value investing, the Benjamin Graham Number offers a simplified yet potent valuation approach. It combines earnings per share (EPS) and book value per share (BVPS) to calculate a stock's intrinsic value. This method provides a quick assessment of stock value, enforcing Graham's principles of investing with a margin of safety.

Earnings Power Value (EPV) estimates the value of a firm based on its current earning power, disregarding future growth expectations. This approach is useful for investors focused on the company's existing operations' profitability, offering a conservative estimate of value that emphasizes sustainable earnings.

3.3 Model Development and Evaluation

Utilizing Python and R, several machine learning algorithms were explored, including neural networks, random forest regression, and simpler linear models (OLS and Lasso). The models were trained on principal components derived from the PCA, optimizing for both predictive accuracy and computational efficiency.

The data set was split into training (90%) and testing (10%) sets. To avoid overfitting, K-fold cross-validation (K=5) was used, employing metrics like R^2, negative MAE, and negative RMSE. The effectiveness of the model's signals was validated by comparing the forecasted and actual year-ahead stock price trajectories. The returns were compared with the return of a reference index over the same period. Success was defined as the return of a model's predictions outperforming the index return. The estimated intrinsic value was reduced by a Margin of Safety (MOS) of 20%. If the resulting valuation was below the current value, the models emitted a ”Buy” signal.

4 Summary of Results and Conclusion

This study's investigation into the integration of machine learning with traditional equity valuation methods unveils insightful outcomes. Noteworthy conclusions are:

- **Model Performance:** Among the models tested, Neural Networks (NN) demonstrated impressive R^2 values, indicating strong predictive capabilities. However, they fell short of accurately generating actionable ”Buy” signals. In contrast, Random Forest (RF) models not only showed high precision in predicting profitable ”Buy” signals but also consistently outperformed the benchmark in terms of returns as detailed in Table 1 in the Appendix.
- **Investment Signals:** With a Margin of Safety (MOS) set at 20%, all models demonstrated the ability to identify lucrative ”Buy” signals, with the RF models surpassing the benchmark by delivering annual returns exceeding 17%.
- **Statistical Significance:** The superiority of RF models over the benchmark was further corroborated by a one-sided t-test, yielding a p-value of 0.03.
- **Data Quality and Feature Engineering:** The importance of data quality and strategic feature engineering was pivotal in model performance. Models using PCA showed markedly better results, highlighting efficient data preparation's role in success. Significant predictors for principal components are detailed in Figures 1 and 2 in the Appendix.

In conclusion, this research provides a foundation for data-driven decision-making in finance, offering insights that can guide investors toward maximizing returns while mitigating market risks. As financial markets evolve, these findings are poised to significantly influence investor strategies in an increasingly data-centric world.

Author Contributions

The author of this paper was responsible for the conception and design of the study, data collection and analysis, and drafting of the manuscript. The author also contributed to the interpretation of data, revising the manuscript critically for important intellectual content, and gave final approval of the version to be published.

URM Statement

The author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Dave Ahern. Explaining the dcf valuation model with a simple example. 2020. URL https://einvestingforbeginners.com/dcf-valuation.

Amir Amel-Zadeh, Jan-Peter Calliess, Daniel Kaiser, and Stephen Roberts. Machine learning-based financial statement analysis. University of Oxford, 2020. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3520684.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep Learning in Asset Pricing, April 2019. URL https://papers.ssrn.com/abstract=3350138.

FMP. Dcf calculations - financialmodelingprep. 2023. URL https://financialmodelingprep.com/dcf-formula.

Benjamin Graham and David Dodd. *Security Analysis*. McGraw-Hill Education - Europe, 1940. ISBN 978-0071412285.

Corporate Finance Institute. Earnings power value. 2023. URL `https://corporatefinanceinstitute.com/resources/valuation/earnings-power-value/`.

Investopedia. Graham number: Definition, formula, example, and limitations. 2023. URL `https://www.investopedia.com/terms/g/graham-number.asp`.

Nikola Milosevic. Equity forecast: Predicting long term stock price movement using machine learning. *ArXiv*, abs/1603.00751, 2016.

A APPENDIX

A.1 FORMULAE FOR THE INTRINSIC VALUE CALCULATION

This section delineates the formulae employed to calculate a company's intrinsic value. These equations utilize a set of fixed inputs—predominantly historical financial data—to estimate the fair value of stocks. The resultant dependent variable, y, is the mean of the outcomes from these formulae, ensuring a balanced assessment of each stock's intrinsic worth. These formulae, central to our methodology, are elucidated below for reproducibility.

A.1.1 DISCOUNTED CASH FLOW (DCF)

The DCF method, highly regarded among value investors, estimates a company's worth based on the present value of its anticipated future cash flows. This method's core advantage is its emphasis on cash flows—a less manipulable metric than earnings—reflecting the genuine economic value to investors. However, DCF's reliability hinges on several assumptions, including the cash flow growth rate, discount rate, and terminal rate, underscoring the importance of judicious parameter selection. The DCF yields a value that should be lower than the asset's current market value for it to be considered a prospective investment (Ahern, 2020).

Projecting future cash flows often involves extending historical trends; however, the absence of an ”ideal” growth rate must be acknowledged due to the inherent uncertainty of future financial landscapes (Ahern, 2020).

The discount rate is calculated using the weighted average cost of capital (WACC). The WACC is in its essence the weighted average of a company's cost of debt and cost of equity (Ahern, 2020).

The needed formulae to calculate the DCF include each of the following:

1) Market Capitalization:

$$\text{Market Cap} = \text{Weighted Average Shares Outstanding Diluted} \times \text{Stock Price}, \tag{1}$$

where ”Weighted Average Shares Outstanding Diluted” refers to the number of outstanding shares of a company, including any potential shares that could be issued from options or convertible securities (FMP, 2023).

2) Enterprise Value Net Borrowings (NB):

$$\text{Enterprise Value NB} = \text{Market Cap} + \text{Long Term Debt} + \text{Short Term Debt}, \tag{2}$$

where ”Long Term Debt” refers to the total amount of debt that a company has outstanding that is due in more than one year. ”Short Term Debt” refers to the total amount of debt that a company has outstanding that is due in less than one year. ”Enterprise Value NB” is a measure of a company's total value, including both its equity and all its net borrowings/debts (FMP, 2023).

3) Equity Value:

$$\text{Equity Value} = \text{Enterprise Value NB} - \text{Net Debt} \tag{3}$$

In this formula, ”Net Debt” refers to the total amount of debt that a company has outstanding, minus any cash and cash equivalents that the company holds. ”Equity Value” is a measure of the total value of a company's equity, after accounting for its debt (FMP, 2023).

4) DCF:

$$\text{DCF} = \frac{\text{Equity Value}}{\text{Weighted Average Shares Outstanding Diluted}}, \tag{4}$$

which results in the first method of valuing a company and which is based on its expected future cash flows (FMP, 2023).

A.1.2 Benjamin Graham Number

The Graham Number is another method to assess the intrinsic value of a stock. The following is the original Graham Formula from ”Security Analysis” (Graham & Dodd, 1940):

$$\text{V} = \frac{\text{EPS} \times (8.5 + 2g) \times (4.4)}{\text{YTM}}, \tag{5}$$

where V is the intrinsic value, EPS the trailing twelve-month earnings per share, 8.5 the *Price-to-Earnings ratio* (*PE ratio*) of a no-growth investment and g being the growth rate for the next seven to ten years.

As the Graham Formula is based on a lot of assumptions about the PE and g and might need adjustments for it to better fit current market circumstances, it has been simplified to the Graham Number, which is given by

$$\text{Graham Number} = \sqrt{(\text{EPS}) \times (\text{BVPS}) \times 22.5}. \tag{6}$$

EPS is calculated by dividing a company's net income by the amount of its common stocks outstanding. The percentage of equity available to common shareholders divided by the total number of outstanding shares is called book value per share ($BVPS$). This figure calculates a company's book value per share and serves as a minimum measure of its equity. It serves as a general test when looking for companies that are currently trading at a good price. Even so, many key elements believed to make a sound investment, such as management quality, significant shareholders, industry characteristics, and the competitive environment, are not considered when calculating the Graham number (Investopedia, 2023).

A.1.3 Earnings Power Value (EPV)

EPV is a valuation method that estimates the present value of a company's future earnings based on its current earnings power, or the ability to generate consistent earnings over the long term.

To calculate EPV, several steps are needed, which involve estimating the company's sustainable earnings, adjusting for non-recurring items, and applying a suitable discount rate to calculate the present value of those earnings.

1) EBIT Margin Averaging

The Earnings Before Interest and Tax (EBIT) margins from the past five years are averaged to mitigate annual fluctuations and ensure a representative profit margin reflective of varying economic cycles.

2) Normalized EBIT Calculation

A' normalized' EBIT is ascertained by multiplying the average EBIT margin with the current sales, which projects future earnings while excluding anomalies and one-off events..

$$\text{Normalized EBIT} = \text{Current Sales} \times \text{Avg EBIT margin}, \tag{7}$$

The normalized EBIT is adjusted for taxes to calculate the after-tax operational earnings, representing the sustainable earning power of the company.

$$\text{After Tax Normalized EBIT} = \text{Normalized EBIT} \times (1 - \text{Effective Tax Rate}) \tag{8}$$

3) Depreciation Adjustment

A tax-adjusted depreciation value is computed to reflect the tax-saving benefits of depreciation.

$$\text{Adjusted Depreciation} = (0.5 \times \text{Effective Tax Rate}) \times \text{Average Depreciation (5 years)} \tag{9}$$

$$\text{Normalized profit} = \text{After Tax Normalized EBIT} + \text{Adjusted Depreciation} \tag{10}$$

The Adjusted Depreciation is calculated by applying the tax rate to the average depreciation expense over the same period. The factor of 0.5 reflects the tax shield effect of depreciation.

4) Maintenance Capital Expenditures (Capex)

To arrive at the maintenance Capex, the average capital expenditure necessary to sustain current operations is calculated next, factoring in the expected income growth.

$$\text{Maintenance Capex} = \text{Total Capex} \times (1 - \%\ \text{Income Growth Rate}), \tag{11}$$

where Total Capex represents the total capital expenditures of the company and $\%\ \text{Income Growth Rate}$ represents the expected annual percentage increase in income.

5) Gross Earnings Power Value

The company's gross earnings power is evaluated by subtracting the maintenance Capex from the normalized profit and then discounting it at the WACC rate.

$$\text{Adjusted Earnings} = \text{Normalized Profit} - \text{Average Maintenance Capex} \tag{12}$$

$$\text{Gross Earnings Power Value} = \frac{\text{Adjusted Earnings}}{\text{WACC}} \tag{13}$$

The Gross Earnings Power Value represents the present value of the expected future earnings of the company, adjusted for the cost of capital.

6) Earnings Power Value

To the gross EPV, any excess net assets are added and debt is subtracted to derive the net EPV, representing the total value accruing to equity holders.

$$\text{Earnings Power Value} = \text{Gross Earnings Power Value} + \text{Excess Net Assets} - \text{Debt} \tag{14}$$

Finally, the EPV per share is calculated by dividing the net EPV by the total number of outstanding shares, indicating the value attributed to each share based on the firm's earning ability.

$$\text{Earnings Power Value per Share} = \frac{\text{Earnings Power Value}}{\text{Number of Shares Outstanding}} \tag{15}$$

These steps ensure a comprehensive understanding of a firm's earning capacity, offering a pragmatic approach to valuation by focusing on the core profitability of the business (Institute, 2023).

B Model Performance and Benchmark Comparison

Table 1 presents the comparative performance of various predictive models against a benchmark. Each model's name indicates the method and any preprocessing steps taken. The table compares the average annual returns generated by "Buy" signals from different models to the performance of a benchmark index during the same periods. ”ReturnsToAvgBenchmark” shows the models’ performance relative to the average benchmark returns, highlighting the efficacy of each predictive model.

Table 1: Returns of Relevant Models & Benchmark

method	buyReturns[a]	benchmarkReturns[b]	returnsToAvgBenchmark[c]
LASSO + PCA	13.23	10.51	1.78
OLS + PCA	14.90	9.98	3.45
Random Forest + PCA (tuned)	17.73	13.80	6.28
Random Forest + PCA (higher bias)	17.19	14.40	5.74
Neural Network + PCA	11.94	12.76	0.49

[a] Average yearly returns of stocks that emitted ”Buy” signals.
[b] Average yearly returns of the benchmark when a stock emitted a ”Buy” signal.
[c] Average yearly returns of ”Buy” signals deduced by the average yearly return of the benchmark.

Abbreviations and Model Names Explained:

LASSO + PCA: Linear model using Least Absolute Shrinkage and Selection Operator with Principal Component Analysis preprocessing, generating ”Buy” signals based on the model’s predictions.

OLS + PCA: Ordinary Least Squares linear regression model with PCA preprocessing, used to generate ”Buy” signals.

Random Forest + PCA (tuned): Tuned Random Forest model utilizing PCA for dimensionality reduction, issuing ”Buy” signals. This model showed the highest outperformance relative to the benchmark.

Random Forest + PCA (higher bias): A variation of the Random Forest model with PCA, adjusted for a higher bias in its predictions, also generating ”Buy” signals.

Neural Network + PCA: Neural Network model employing PCA for data preprocessing, designed to issue ”Buy” signals based on its output.

C Influential Predictors in Principle Components Analysis

C.1 First Principal Component Analysis

The first principal component (PC1) captures variance through a combination of variables that can be broadly categorized into liquidity, profitability, and tax-related factors. Below a detailed explanation of selected influential variables in PC1 is provided:

- CashAndCashEquivalents: This variable indicates a company’s liquid assets that are readily available for use. A higher value suggests a strong liquidity position, which can positively impact stock valuation.
- Revenue: Represents the total income received from normal business operations. It reflects the company’s capacity to generate sales and maintain operational efficiency.
- DeferredTaxLiabilitiesNonCurrent: This metric includes taxes that are accrued but not due for payment in the current period, which can affect a company’s future cash flows and valuation.
- DividendsPaid: Reflects the total amount returned to shareholders. Regular dividends can signify financial stability, while a high payout may indicate a mature company with limited growth opportunities.
- TaxNormalizedEBIT: Earnings before interest and taxes (EBIT), adjusted for a normalized tax rate, provide an indication of operational profitability, excluding the effects of financial structure and tax strategies.
- OtherNonCurrentInvestments: These may represent long-term strategic investments a company has made, which could bear on future profitability and risk.
- PurchasesOfInvestments: Large purchases of investments might indicate a company’s strategy to allocate surplus cash, potentially impacting its growth trajectory and risk profile.

C.2 Second Principal Component Analysis

The second principal component (PC2) emphasizes growth prospects and return on investment to shareholders, evident through the following variables:

- growthEPS: The growth in Earnings Per Share (EPS) is a direct measure of a company's profitability growth, indicating its potential to increase value over time.
- RevenueGrowthShare: This variable combines revenue growth with the per-share basis, focusing on the company's ability to scale its operations effectively.
- threeYearDividendGrowthShare: This signifies the compound growth rate of dividends over three years. An increasing trend may be interpreted as a signal of confidence by management in the company's future earning potential.
- NetIncomeGrowthShare: Growth in net income adjusted per share can reflect a company's increasing efficiency or market expansion.
- OperatingCashFlowGrowthShare: Cash flows from operations are crucial for sustaining and growing a business. This metric's growth rate is often considered more reliable than earnings growth because it's harder to manipulate with accounting practices.
- ShortTermInvestments: These are assets that can be quickly converted into cash, usually within a year, and reflect a company's tactical financial decisions.

Figures 1 and 2 illustrate the relative contribution of these variables to PC1 and PC2, respectively. These figures enable a comprehensive view of the financial dimensions that are most influential in the machine learning models' assessment of stock valuation, providing a nuanced understanding of the features driving predictive performance.

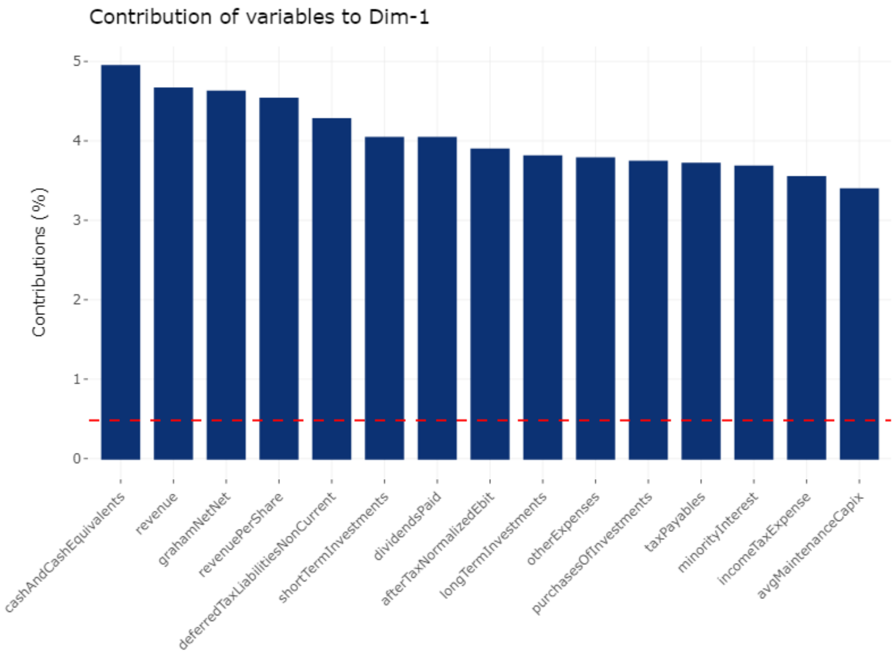

Figure 1: Top 15 Most Influential Predictors on the First Principle Component

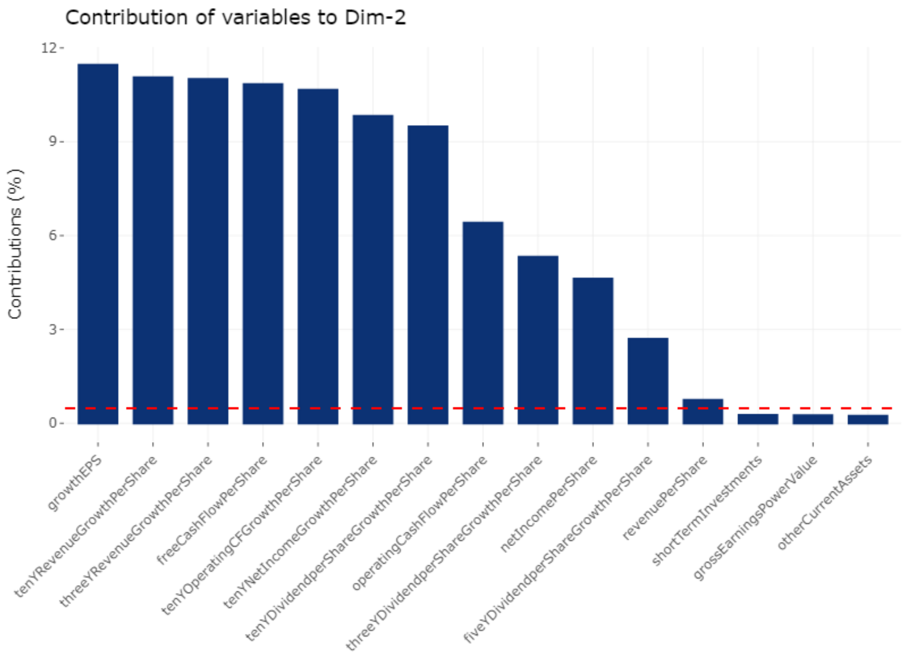

Figure 2: Top 15 Most Influential Predictors on the Second Principle Component