15.433 - Financial Markets – Final Project Report

Developing a Multifactor Model to Identify Winning Growth Stocks

Avijit Srivastava – avijitns@mit.edu - 975316276

Paul O'Connor – paulroc@mit.edu – 987101509

Prapti Kumar – prapti@mit.edu – 952590591

Yash Gupta – ygupta@mit.edu – 938024529

Growth Stocks

Introduction

Growth Stocks are stocks that are expected to grow in the future by an above average rate when compared to the other stocks in the market. They usually trade at a market value higher than their intrinsic value due to their future growth potential; hence, they have a low Book Value/Market Value (B/M) ratio and a high Price/Earnings (PE) ratio. Usually, growth stocks pay zero to very low dividends, for the usual belief is that they can generate a higher return through investing the money back. Investors generally buy growth stocks in the expectation of capital gains, when they are more interested in future years' earnings than in next year's dividends. The idea behind the growth strategy is the efficient market hypothesis which states that the current stock price reflects all the information available about the firm and, therefore, the current price is most reasonable at that point of time (An et al.).

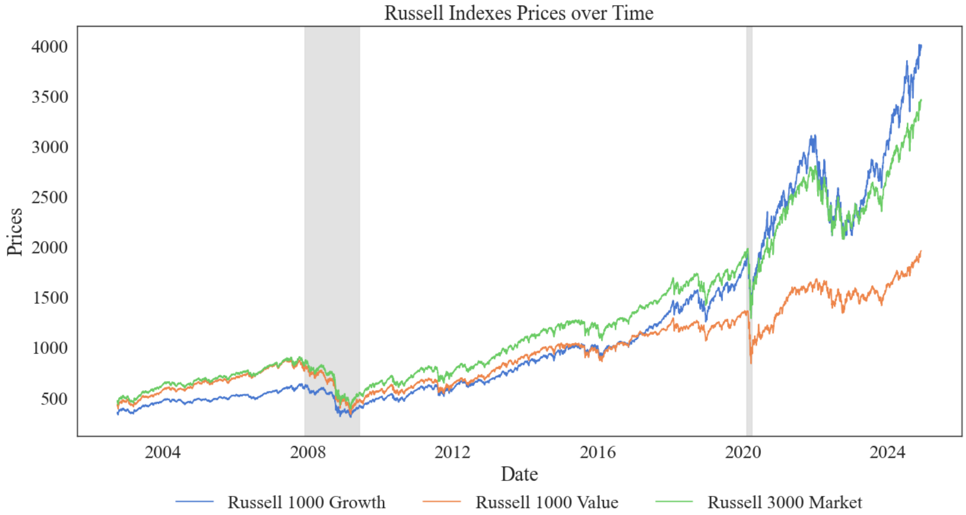
Difference between Growth and Value Stocks

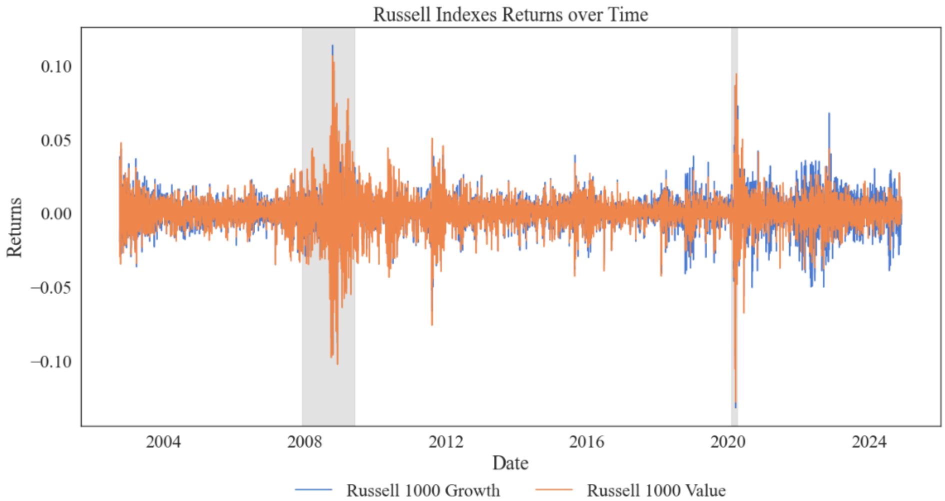
Growth and value investment strategies are amongst the most popular investment strategies used by professional investors. Both strategies suggest different views in the investment community, and there have often been questions on which strategy performs better. Through this project, we aim to highlight both these strategies, the differences between them, and investigate further a model to identify winning growth stocks.

The difference between value stocks and growth stocks arises from the difference in perspectives about the companies. Value stock investors look for stocks that have high value ratios and believe that these stocks have strong fundamentals for book value and earning power but currently are undervalued in the market now (An et al.).

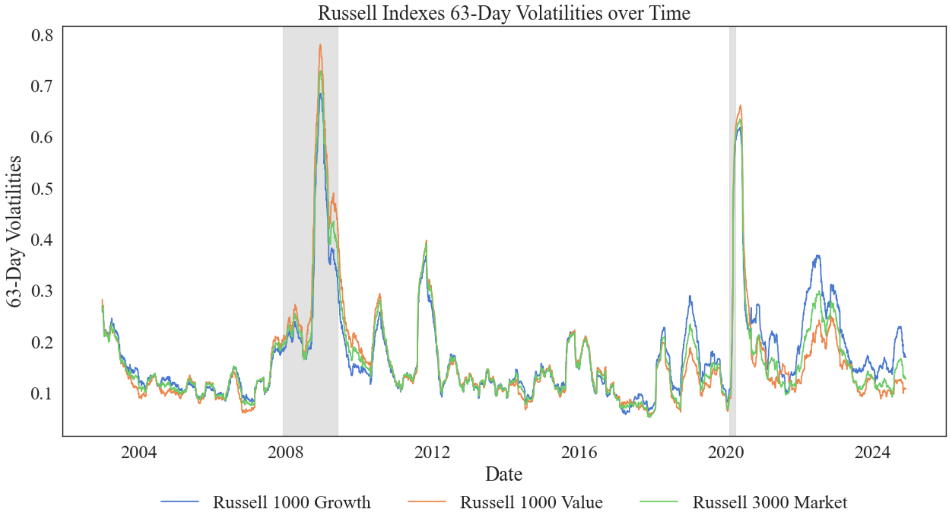
Growth stocks typically grow significantly faster than their counterparts in terms of sales and, above all, profits. They are young, innovative and agile companies that differ in many ways from mature, highly cyclical value stocks. The growth segment is riding a wave of success as disruptive ideas and technologies capture the zeitgeist and offer hope for future growth.

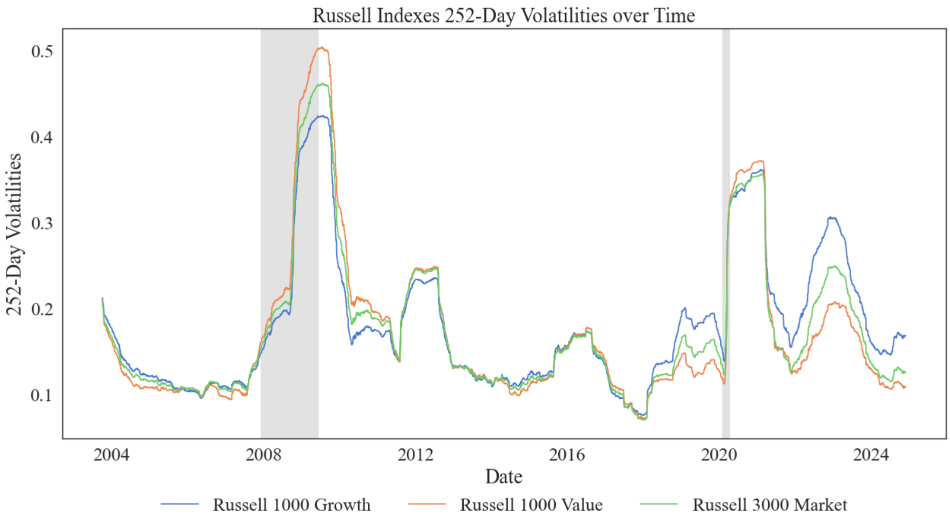
Historically, it is often observed that growth stocks perform well in bullish markets, while value stocks provide security during downturns. In recent years, the outperformance of the growth stock sector versus the value sector is clear. In the plots below, the blue line indicating the Russell 1000 Growth Index has higher cumulative returns and higher daily returns since before the Covid-19 pandemic. The distribution of returns also appears to be similar for the growth stocks having potentially higher skew and more extreme outliers (both positive and negative).





Growth Stocks are considered more volatile and involve a higher level of risk and returns. In contrast, value stocks deliver a steadier performance, offering stability to investors in stressful environments. This perspective is not as evident in the data, as can be seen in the plots below with the Value index having the highest realized rolling volatility for a large proportion of the period. However, in recent years the volatility of the Growth index has been higher. This of course takes both up and downside risk into account.





Value stocks are perceived to have lower P/E ratios since that might indicate that the stock is undervalued compared to its earning potential. These stocks are usually associated with low P/B, low price/cash flow, and a higher and more consistent dividend yield in contrast to growth stocks.

Investors buy stocks for two reasons: the expectation of capital gains and cash dividends. For growth stocks, the first reason dominates; that is, they are more interested in the future growth of earnings (and hence stock price) than in the next year's dividends. On the other hand, they buy value stocks primarily for their cash dividends. This distinction makes sense since the return on equity for growth stocks exceeds their cost of capital ($ROE > r_e$).
From a Corporate Finance lens, the price of a stock today is

$$P_0 = PVGO + \frac{E_1}{r_e};$$

That is, it can be decomposed into a constant perpetuity of next year's earnings and the present value of growth opportunities (PVGO). For the growth stocks, the PVGO is positive and hence

$$\frac{P_0}{E_1} = \frac{PVGO}{E_1} + \frac{1}{r_e} > \frac{1}{r_e}$$

which means that ceteris paribus the P/E ratio of growth stocks should be higher compared to similar stocks.

Since for growth stocks, it makes more sense for the firm to re-invest its earnings in positive NPV projects, creating value for the shareholders. On the other hand, for value stocks, it makes more sense to distribute the earnings back to the shareholders in the form of cash dividends who can then find better NPV opportunities themselves. This choice is, in fact, observed for real-world firms. For example, many big-tech firms, including Google, have been reinvesting their profits for the bulk of their starting years due to the immense growth (positive NPV) opportunities they have. In contrast, AT&T, with limited growth opportunities, is known for its high dividend yield.

Risk and Reward Characteristics

Investing in growth stocks offers a compelling risk-reward profile that attracts investors seeking above-market returns. The key appeal lies in their potential to deliver significant future earnings growth, which often justifies their higher valuation, and the premium investors pay for them. However, this also makes growth stocks highly volatile, as their prices are heavily influenced by market sentiment and forecasts. Any shift in expectations and/or increased competition can lead to price fluctuations. Investors may overreact or underreact to both good and bad news making the volatility worse. Over time, the initially high Price-to-Book (P/B) ratios of growth stocks often decline as competition increases and market sentiment changes (Fama and French, 2007). While some companies manage to sustain their high P/B ratios, they are the rare exceptions.

Challenges in Modeling Growth Stocks

- **High Valuation:** Growth stocks are characterized by high valuation ratios based on forecasts of future performance. This creates a significant challenge in determining whether these valuations are justified or based on overly optimistic or erroneous forecasting, making accurate modeling complex and prone to error and subjectivity.

- **Cyclical Performance:** The performance of growth and value stocks tends to oscillate with business cycles, making it difficult to make general claims like "growth stocks outperform value stocks in the long run." During economic booms (e.g., the 2008–21 bull market or the late '90s dotcom rally), growth stocks often exhibit sustained positive performance. However, in times of economic strain or crises, their performance reverses as investors refocus on fundamentals like earnings and dividends. Selecting the appropriate economic regime is crucial, as factor influences can vary significantly with changes in the business cycle. Hartford notes that past success of any particular investing style does not guarantee future returns—for instance, big tech stocks may cease to soar indefinitely. Figure 1 illustrates this cyclical behavior, emphasizing the importance of selecting the appropriate regime or business cycle for modeling.

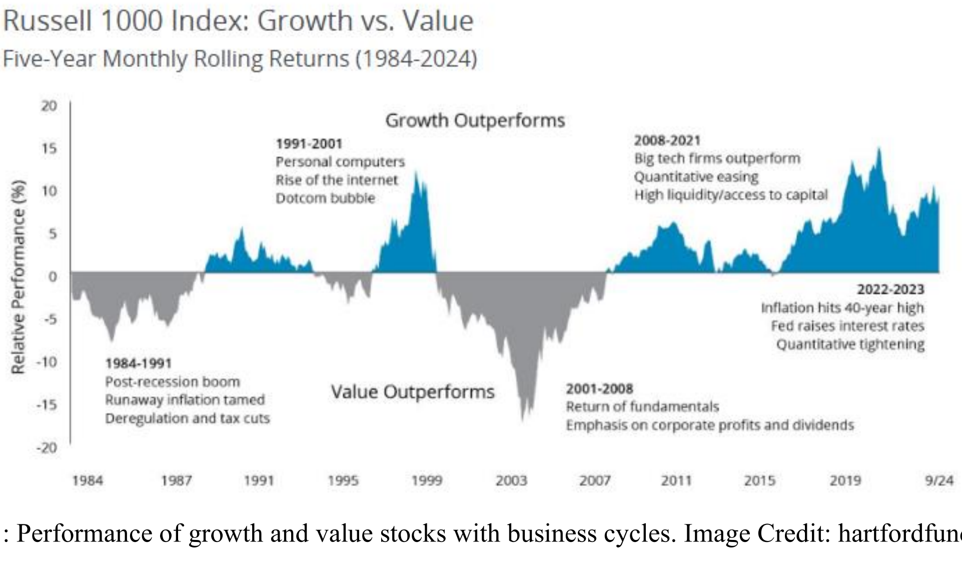


Figure 1: Performance of growth and value stocks with business cycles. Image Credit: hartfordfunds.com

- **Sensitivity to Economic Changes:** Growth stocks, particularly in their early stages, are often capital-intensive and thus highly sensitive to macroeconomic factors such as interest rates and inflation. For example, the post-GFC outperformance of growth stocks might have been driven by low interest rates and a strong economy (Bevanda et al., 2021). Any shifts in these conditions can significantly impact their valuation and returns.

- **Competition and Technological Innovation:** The performance of growth stocks is often influenced by unpredictable factors like market competition, technological breakthroughs, and regulatory changes. These uncertainties make it challenging to accurately incorporate such variables into financial models, increasing the difficulty of forecasting.

- **Management and Talent**: The success of growth stocks might be due to the quality of the company's management and talent. However, evaluating these qualitative factors for individual stocks is extremely laborious and difficult to standardize, hence adding to the complexity of modeling.

Model Development - Data and Methodology

Our models used data on a range of Russell 1000 Growth Index members from 2010 to 2024. We had daily price data for each as well as annual financial metrics and ratios including Revenue, EBIT, Earnings Per Share (EPS), Debt-to-Equity (DE) Ratio, Return on Equity (ROE), Return on Assets (ROA), Analyst Ratings, Return on Invested Capital and EBIT Margin.

Factor Sorting and Modelling

At each date, stocks were sorted into portfolios based on the selected financial metrics: Analyst Ratings, ROIC, and EBIT Margin. For instance, using the Analyst Ratings metric, Rating 1 represents the portfolio comprising stocks with the worst analyst ratings for a given year and Rating 5 the stocks with the best ratings, with portfolio composition rebalanced annually. Equal-weighted returns were calculated for each portfolio and tracked over time. The monthly returns of these sorted portfolios were then regressed against the Fama-French 3-Factor (FF3) model to determine whether the selected factors offered any additional explanatory power beyond what the FF3 model captures. This process was repeated independently for each factor to ensure a consistent evaluation across all metrics.

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
RATING-1.0	**-0.007**	0.000	**1.171**	0.000	**0.568**	0.000	**0.399**	0.000	0.921
RATING-2.0	**-0.004**	0.010	**1.343**	0.000	**0.499**	0.000	**0.242**	0.006	0.910
RATING-3.0	**-0.004**	0.009	**1.407**	0.000	**0.479**	0.000	0.129	0.078	0.918
RATING-4.0	**-0.005**	0.006	**1.479**	0.000	**0.427**	0.000	0.006	0.937	0.920
RATING-5.0	-0.002	0.385	**1.416**	0.000	**0.490**	0.000	-0.030	0.742	0.896

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
EPS-1.0	**-0.007**	0.006	**1.318**	0.000	**1.133**	0.000	-0.032	0.776	0.855
EPS-2.0	-0.003	0.104	**1.310**	0.000	**0.529**	0.000	0.055	0.566	0.903
EPS-3.0	-0.003	0.020	**1.281**	0.000	**0.289**	0.000	**0.218**	0.001	0.920
EPS-4.0	**-0.004**	0.002	**1.301**	0.000	**0.241**	0.000	**0.248**	0.000	0.923
EPS-5.0	**-0.004**	0.001	**1.354**	0.000	**0.213**	0.001	**0.290**	0.000	0.920

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
PE-1.0	**-0.008**	0.000	**1.322**	0.000	**0.510**	0.000	**0.709**	0.000	0.904
PE-2.0	-0.003	0.072	**1.253**	0.000	**0.401**	0.000	**0.438**	0.000	0.913
PE-3.0	-0.003	0.058	**1.303**	0.000	**0.271**	0.000	**0.289**	0.000	0.920
PE-4.0	-0.002	0.373	**1.382**	0.000	**0.320**	0.000	0.021	0.779	0.905
PE-5.0	**-0.005**	0.005	**1.442**	0.000	**0.550**	0.000	**-0.279**	0.000	0.893

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
DE-1.0	-0.001	0.569	**1.194**	0.000	**0.633**	0.000	**-0.318**	0.000	0.885
DE-2.0	**-0.004**	0.003	**1.232**	0.000	**0.411**	0.000	**0.215**	0.000	0.924
DE-3.0	**-0.004**	0.010	**1.289**	0.000	**0.443**	0.000	**0.275**	0.001	0.916
DE-4.0	**-0.006**	0.001	**1.367**	0.000	**0.437**	0.000	**0.232**	0.003	0.909
DE-5.0	**-0.008**	0.000	**1.441**	0.000	**0.525**	0.000	**0.349**	0.000	0.899

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
ROE-1.0	**-0.008**	0.000	**1.378**	0.000	1.044	0.000	0.035	0.778	0.863
ROE-2.0	**-0.005**	0.000	**1.189**	0.000	0.443	0.000	**0.272**	0.001	0.915
ROE-3.0	**-0.004**	0.003	**1.239**	0.000	0.298	0.000	**0.243**	0.000	0.921
ROE-4.0	-0.002	0.079	**1.369**	0.000	0.279	0.000	0.122	0.025	0.919
ROE-5.0	-0.003	0.071	**1.508**	0.000	0.239	0.000	0.185	0.011	0.906

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
ROA-1.0	**-0.008**	0.001	**1.349**	0.000	**1.150**	0.000	-0.065	0.537	0.861
ROA-2.0	-0.004	0.023	**1.257**	0.000	**0.450**	0.000	**0.400**	0.000	0.892
ROA-3.0	**-0.004**	0.001	**1.250**	0.000	**0.323**	0.000	**0.295**	0.000	0.924
ROA-4.0	-0.003	0.098	**1.332**	0.000	**0.278**	0.000	0.183	0.012	0.906
ROA-5.0	-0.002	0.144	**1.444**	0.000	**0.205**	0.008	-0.099	0.083	0.901

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
ROIC-1.0	**-0.007**	0.002	**1.326**	0.000	**1.136**	0.000	-0.084	0.388	0.860
ROIC-2.0	**-0.006**	0.001	**1.243**	0.000	**0.418**	0.000	**0.347**	0.001	0.898
ROIC-3.0	**-0.005**	0.001	**1.251**	0.000	**0.332**	0.000	**0.308**	0.000	0.918
ROIC-4.0	-0.002	0.174	**1.364**	0.000	**0.287**	0.000	**0.224**	0.005	0.915
ROIC-5.0	-0.001	0.394	**1.451**	0.000	**0.241**	0.001	-0.084	0.138	0.903

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
EBIT Margin-1.0	-0.013	0.486	**2.342**	0.000	0.902	0.431	-1.011	0.065	0.279
EBIT Margin-2.0	0.009	0.215	**2.166**	0.000	0.270	0.462	-0.134	0.374	0.441
EBIT Margin-3.0	-0.001	0.958	**2.563**	0.000	0.935	0.049	-0.430	0.261	0.456
EBIT Margin-4.0	0.013	0.193	**1.911**	0.000	0.111	0.727	-0.162	0.460	0.602
EBIT Margin-5.0	0.004	0.674	**2.247**	0.000	0.144	0.632	-0.036	0.858	0.775

When analyzing the regression results the high R-squared values indicate that the Fama-French 3 factor model explains most of the variation in the sorted portfolios. With the exception of EBIT margin which has lower values for some of the portfolios. Coefficients which are significant at the 1% level are bolded. The market factor is significant in all models. Overall, the results point to limited new explanatory power in our potential factors. We also attempted to sort by multiple factors which had slightly lower R-squared values.

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
ROIC-PE-1-1	**-0.010**	0.003	**1.638**	0.000	**0.963**	0.000	**0.953**	0.000	0.863
ROIC-PE-2-2	-0.002	0.365	**1.784**	0.000	**0.337**	0.000	**0.421**	0.000	0.897
ROIC-PE-3-3	0.002	0.597	**3.188**	0.000	0.533	0.015	**-1.306**	0.000	0.857

Portfolio	Intercept Coefficient	Intercept P-Value	Mkt-RF Coefficient	Mkt-RF P-Value	SMB Coefficient	SMB P-Value	HML Coefficient	HML P-Value	Adj R-Squared
RATING-EBIT Margin-1-1	-0.008	0.639	**2.939**	0.000	**2.067**	0.009	-0.132	0.801	0.651

RATING-EBIT Margin-2-2	0.006	0.438	**3.972**	0.000	0.317	0.514	0.137	0.550	0.786
RATING-EBIT Margin-3-3	0.011	0.497	**3.706**	0.000	-0.641	0.250	-0.440	0.287	0.717

Ultimately, we decided to focus on four key factors to evaluate growth stocks: Analyst Ratings, Return on Invested Capital (ROIC), EBIT Margin, and the Price/Earnings (PE) Ratio. While some of these factors had high R-squared values when regressed against the FF3 model we believed in the economic rationale for their selection. These factors were selected for their ability to capture forward-looking expectations, management efficiency and operational profitability—critical aspects of identifying winning growth stocks.

Analyst Ratings provide a market-driven, forward-looking perspective by reflecting expert forecasts on a company's future performance which is utilized in our project to get an idea of market expectation for the growth trajectory and future earning of a stock, making it an essential tool for identifying potential winners.

ROIC measures how effectively a firm utilizes its invested capital to generate returns, providing valuable insights into management efficiency—a critical foundation for a company's success and its potential for sustained growth. It is especially critical for growth stocks to reinvest earnings to expand. This also highlights the firm's ability to allocate resources effectively, which is vital for sustaining long-term growth and creating shareholder value.

EBIT Margin evaluates operational efficiency, which is especially relevant for growth companies navigating competitive and resource-intensive environments. It provides a clear picture of whether a company can scale its operations while maintaining profitability.

Finally, the PE ratio reflects market confidence in a stock's ability to generate future earnings around which the entire idea of a growth stock revolves. The PE ratio also highlights the premium investors are willing to pay for anticipated future earnings, serving as a central metric for assessing whether a stock aligns with growth expectations. Together, these factors create a balanced approach for analyzing growth stocks.

Our model takes the following form:

$$r_{i.t} = \alpha + \beta_1 * f_{rating} + \beta_2 * f_{EBIT\ margin} + \beta_3 * f_{price-earnings} + \beta_4 * f_{roic} + \epsilon_{i,t}$$

where each factor represents the monthly returns from going long the highest quintile of that ratio (e.g. Rating 5) and short the lowest quintile (e.g. Rating 1), $r_{i,t}$ is the month t excess return from growth stock i. We fit the model to all the growth stocks in our data set and it exhibited similarly poor explanatory power to the Fama-French 3-Factor model, as can be seen by the average statistics across all the growth stocks in the Russell 1000 Growth Index. Both models struggled to explain the cross section of returns with low R-Squared values and high P-Values.

Fama-French 3-Factor	
R-squared	0.005
Constant P-Value	0.461
Mkt-RF P-Value	0.501
SMB P-Value	0.413
HML P-Value	0.370

Winning Growth Factor Model	
R-squared	0.008
Constant P-Value	0.487
Rating	0.472
PE	0.402
ROIC	0.493
EBIT Margin	0.374

Classification Model

A classification model was developed based on the factors - Debt to Equity Ratio (DE), EBIT, EPS, Equity value, PE ratio, Analyst Rating, ROA, ROE, ROIC and Shares outstanding. Data was collected for each ticker symbol on a yearly basis, and considering a holding period of one year, the model would predict if a stock would classify as a high-potential growth stock or not (binary classification, 1 corresponding to a high-potential growth stock).

Data Preprocessing

After ensuring that none of the features had any missing values, we inspected their distributions to gauge whether any transformation is required before imputation into the model. Unfortunately, they were found to have highly skewed distributions and significant outliers, signaling the risk of high levering effect on linear models such as logistic regression. Figure 2 below shows the highly right-skewed distribution of Equity and highly left-skewed distribution of ROA.

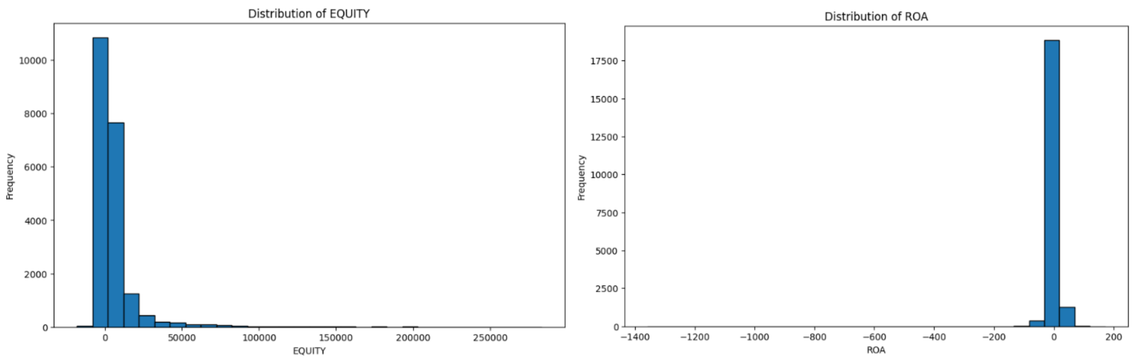


Figure 2: Distribution of values assumed by the features EQUITY and ROA.

As is the standard in the literature, to alleviate the issue we decided to use log-transformations. Specifically, we used the following to also preserve sign information:

$$Transformed\ Feature = sgn(Orig\ Feature) \times \ln(Orig\ Feature + 10^{-5})$$

A small constant, 10-5, was added to avoid invalid (near zero) inputs and exploding values associated with the log function.

The effect of the transformation is illustrated for the features above in Figure 3 below.

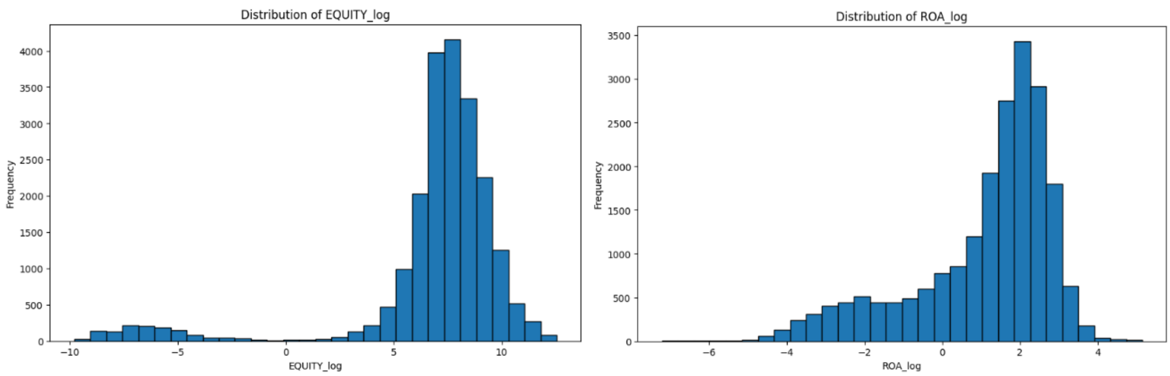


Figure 3: Distribution of log-transformed features (Equity and ROA) having a lower range and no significant outliers compared to the original.

The log-transformation also allows us to use the Standard Scaling operation in the model pipeline. The standard scaling allows for faster convergence and further mitigates the effect of extreme values.

For defining the labels for high potential vs other stocks (1 vs. 0), we employed the following two thresholding methods:

1. *Relative Benchmark Thresholding:* If the return from the stock is greater than the return from Russell index for the year, classify as 1, else 0. This approach, however, may produce low-quality labels for training since several stocks would be near the boundary (1 vs. 0) and would not quite represent either category.

2. *Volatility-Adjusted Benchmark Thresholding*: With the hyperparameter k,
Label 1 if
$$ret[stock_i] > ret[russell1000] + k \times \sigma[russell1000]$$
Otherwise, label 0 if
$$ret[stock_i] < ret[russell1000] - k \times \sigma[russell1000]$$
Else, the record would be dropped for training/validation purposes (but not for developing the trading strategy later). This ensures that we are including only the instances we are confident about in training.

The hyperparameter k was tuned (choosing values ranging from 0 to 3, inclusive). It was noticed that while higher values of k boosted the in- and out-of-sample performance, they were also associated with increasing shrinking data (since more and more records were dropped). Therefore, we chose k to be 1 to allow for a balance between performance and size of the data (to allow the results to be statistically significant).

Modelling
As Occam's Razor dictates, we started with the simplest classification model - the Logistic Regression, and then moved to the more complex XGBoost and Neural Network models.

Logistic Regression: A linear combination of the factors, called the logit, is expressed as
$$g = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_p \times x_p$$
Then the probability of the label being 1 becomes
$$y = \frac{e^g}{1 + e^g}$$
This simple model is fast to train and infer from as well as less susceptible to overfitting. The model is fitted to minimize the binary cross-entropy loss defined below, where m is the number of training instances.
$$J = \frac{1}{m}\sum_{i=1}^{m}\left(y_i \times \log\left(p(y_i)\right) + (1 - y_i) \times \log\left(1 - p(y_i)\right)\right)$$

XGBoost: Decision trees can be combined in a boosting-based ensemble, each successive tree trained to predict the residuals of the prediction by the first. This more complex model produced significant overfitting, hence we tuned it using Randomized Grid Search to set some of the hyperparameters to produce higher out-of-sample performance. Specifically, the number of features per tree, the maximum depth of each tree, the learning rate, the subsample of data to be used, and number of trees in the ensemble. Randomized Search was preferred over the regular Grid Search for computational reasons. The tuning helped mitigate the overfitting as well as boosted the out-of-sample performance.

Neural Network: Figure 4 below shows the topology of the fully-connected network we used. Dropout and Batch normalization layers were added to mitigate the issue of overfitting associated with complex models. Additionally, He-Normal initialization was used for faster convergence to the results. Mini-batch gradient descent with Adam optimizer was used for fitting the model.

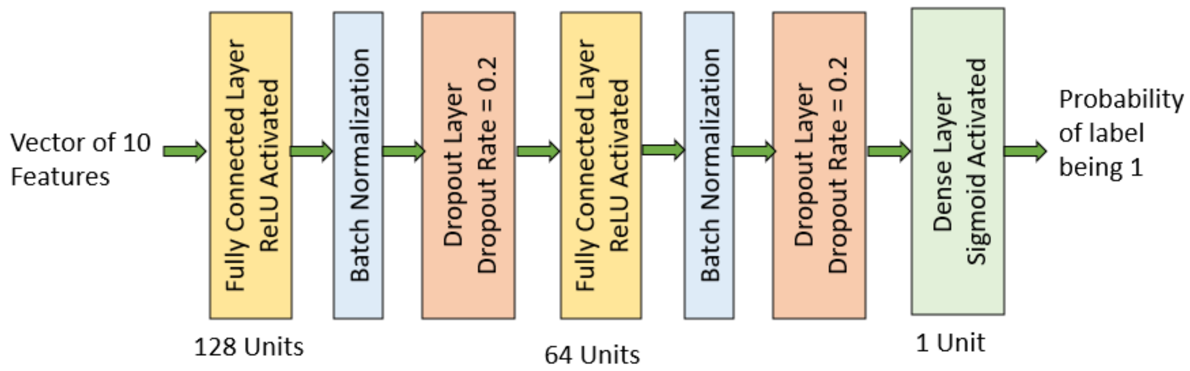


Figure 4: The Neural Network Architecture

Model Evaluation Classification Model

Data Analysis

We split the dataset into training and testing partitions. We consider the period between 2010 to 2017 to be our training dataset, and the period between 2018 to 2019 to be our test data for the following performance analysis in terms of standard metrics for classification for the three models considered. Thereafter, we consider periods beyond 2019 (up to 2023) for testing the performance in terms of log returns of the stocks selected by the model against the Russell-1000.

1. Relative Benchmark Thresholding (RBT)

On training the model based on a simple labelling method, i.e., just comparing the stock return to the russell index annual return to classify as 1 or 0 (high growth potential vs low), we fit four different models on our dataset - Logistic Regression, XGBoost Classifier, Neural Network and a tuned XGBoost we find that we achieve a training accuracy of 58.68% and a test accuracy of 64.93%. We further fit a XGBoost Classifier on the dataset, achieving an in sample accuracy of 93.56% and an out of sample accuracy of 61.3%, indicating overfitting on the data.

Results:

Metric	**Logistic Regression**	**XGBoost**	**Neural network**	**Tuned XGBoost**
In Sample				
Accuracy	58.68%	93.56%	60.81%	63.71%
Precision	57.25%	93.35%	59.08%	63.40%
Recall	57.19%	93.32%	61.47%	58.93%
F1-Score	57.22%	93.34%	60.25%	61.08%
Out of Sample				
Accuracy	64.93%	61.26%	62.14%	65.80%
Precision	61.38%	58.03%	58.36%	63.59%
Recall	63.30%	55.91%	60.70%	59.39%
F1-Score	62.33%	56.95%	59.51%	61.42%

2. Volatility-Adjusted Benchmark Thresholding (VBT)

We use a more conservative labelling, where we incorporate a margin k scaled by the standard deviation of the Russell 1000 Growth Index to account for volatility in the classification decision.

We see that XGBoost's out-of-sample accuracy improves from 61.26% (RBT) to 71.08% (VBT), demonstrating better generalization. The Neural network out-of-sample accuracy increases to 67.52% in VBT, but precision decreases (53.51%), suggesting higher false positive rates.

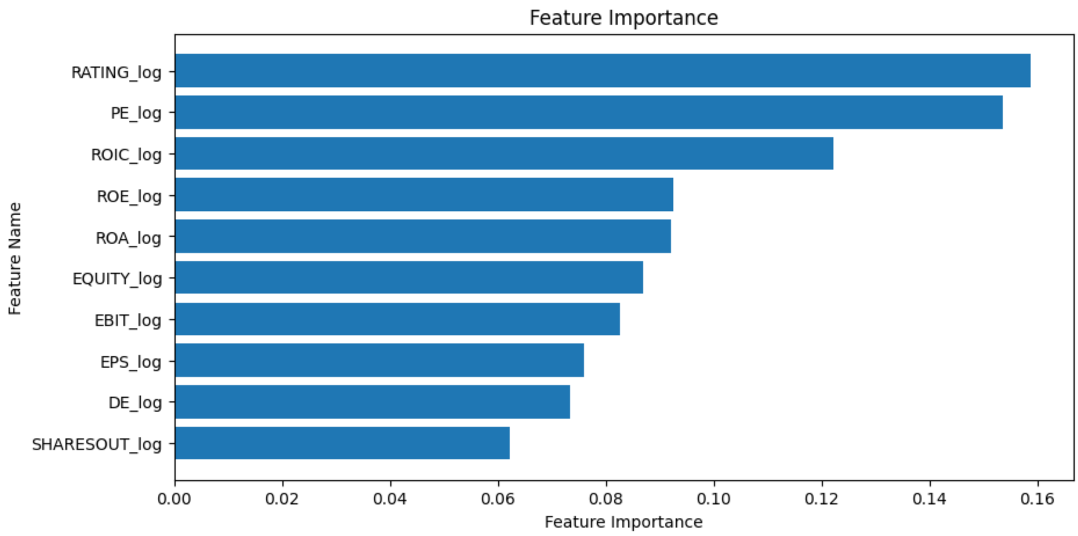
Results:

Metric	Logistic Regression	XGBoost	Neural network	Tuned XGBoost
In Sample				
Accuracy	64.40%	99.29%	62.41%	71.61%
Precision	62.42%	99.07%	57.30%	70.41%
Recall	59.21%	99.42%	75.69%	67.33%
F1-Score	60.77%	99.24%	65.22%	68.84%
Out of Sample				
Accuracy	70.44%	71.08%	67.52%	77.46%
Precision	57.44%	58.71%	53.51%	66.74%
Recall	69.37%	66.58%	75.19%	74.68%
F1-Score	62.84%	62.39%	62.53%	70.49%

The Relative Importance of Features

We inspect the relative importance of the different features in driving the classification predictions of the model. The tuned XGBoost model, which gives the best out-of-sample performance, is used for this purpose. Figure 5 below illustrates the same as a bar graph.

It is easy to note that while certain features such as Ratings, PE ratio, and ROIC prove of higher importance in driving the predictions, the difference between them and the other features is not that high. This suggests that dropping the features might be a bad idea. In fact, we tested retaining just the top 5 features and observed that the XGBoost model's performance plummeted from 77.5% to just 55.9%. Therefore, we recommend using all 10 factors instead of just 3-5.



A Strategy based on the Classification Model

Instead of going long on the entire Russel growth index, we explore the possibility of holding only those stocks that actually have a high-growth potential (as identified by our classification model). Specifically, we sort the stocks by the probability of their label being a 1 and select the top 10% of the stocks. These stocks are held for 1 year and the returns are recorded. The process is repeated again. We compare the performance of Russell against these top 10% stocks for each 1-year holding period, as illustrated in the plots below.

Figure 6 shows the performance with a logistic regression while Figure 7 shows that with an XGBoost (using the parameters found after tuning earlier). For each of the 1-year holding periods, the model is trained on all the years up to (but not including) that year. It is easy to note that for each of the 6 years studied, the stocks chosen by the model consistently outperform the Russell growth index. Furthermore, the tuned XGBoost model's higher accuracy indeed translates into a superior performance over the logistic regression.

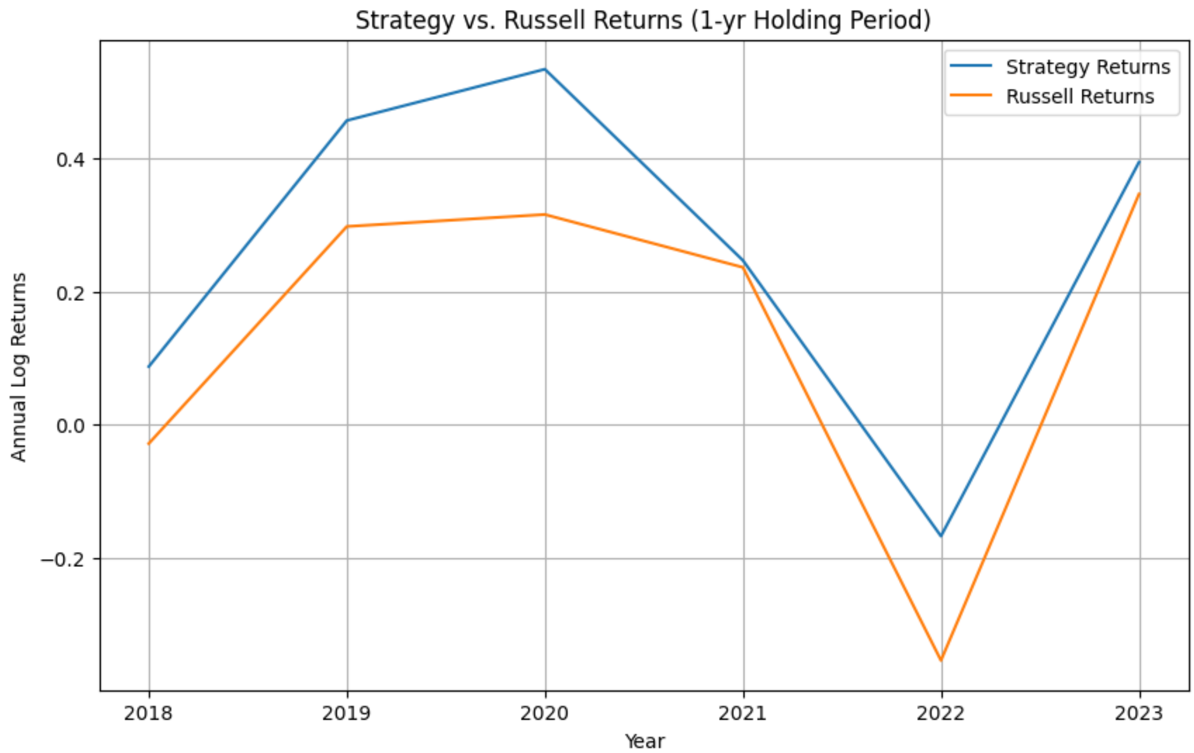


Figure 6: 1-year log returns of the top 10% high-potential growth stocks (as per model predictions) vs. Russel using Logistic Regression

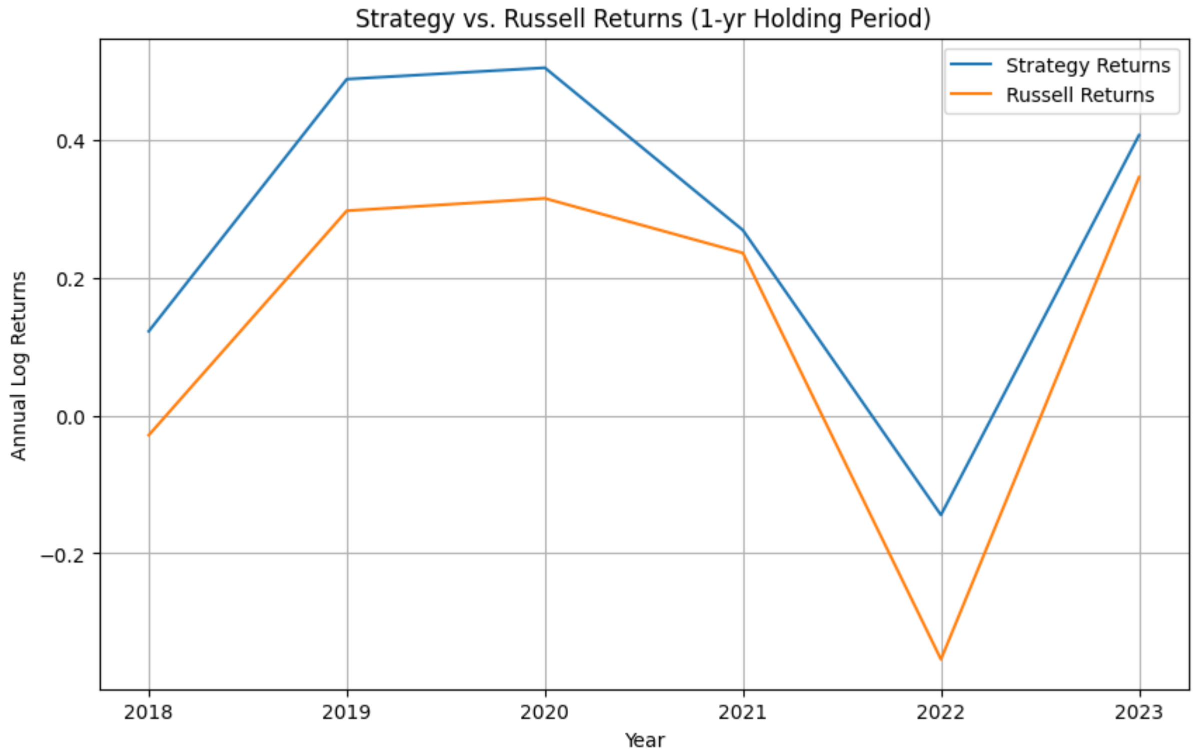


Figure 7: 1-year log returns of the top 10% high-potential growth stocks (as per model predictions) vs. Russell using Tuned XGBoost.

Findings and Recommendations

The results reflect the Efficient Market Hypothesis (EMH), which posits that stock prices incorporate all publicly available information. From our findings, we see that the use of publicly available factors, which are easy to compute and widely available on the internet, result in limited differentiation between stocks because these factors are likely already reflected in stock prices. As a result, their ability to explain cross-sectional stock returns is challenging, especially in highly competitive markets.

However, our classification approach does appear to have improved predictive performance and the ability to identify winning growth stocks. The more advanced machine learning algorithms used are potentially more effective capturing complex trends and in differentiating between high and low performing growth stocks than the linear models.

Overall, our results show the difficulty in predicting the cross-section of returns in modern financial markets. With the wide availability of data and computational resources it is becoming increasingly difficult to find undiscovered alpha generating ideas and models.

References

An, Chongsoo, et al. "Do Value Stocks Outperform Growth Stocks in the U.S. Stock Market?" *Journal of Applied Finance & Banking*, vol. 7, no. 2, 2017, pp. 1792–6599, www.scienpress.com/Upload/JAFB/Vol%207_2_7.pdf.

Bevanda, Lea-Marija, et al. "Performance of Value and Growth Stocks in the Aftermath of the Global Financial Crisis." *Business Systems Research Journal*, vol. 12, no. 2, 1 Dec. 2021, pp. 268–283, https://doi.org/10.2478/bsrj-2021-0032.

Fama, Eugene F., and Kenneth R. French. "The Anatomy of Value and Growth Stock Returns." *SSRN Electronic Journal*, 2007, https://doi.org/10.2139/ssrn.806664. Accessed 18 Sept. 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=806664