Predicting Stock Returns of USA Listed Technology Companies Using Financial Ratios

Yimeng Xiao(yx6df), Xinru Cheng (xc9mb), Xinyue Qiu(xq8zu)

1 Introduction

Investment in financial securities, especially stocks, has become an important aspect of the American economy since the last century. Each day, over 1.5 billion shares of stock are traded on the New York Stock Exchange with millions of people relying on stock trading as their main source of income. At the same time, professional investors and researchers have committed extensive effort to analyze the market in order to gain a competitive advantage over other investors but have not yet been able to find a way to consistently superior returns.

Almost without exception, researchers before the 1980s concluded that the efficient-market hypothesis offers a remarkably good capture of the reality of stock markets. The hypothesis states that all information available to a market is already reflected in stock price, making it impossible to predict changes in stock price without insider information. However, since the 1980s, cracks in this hypothesis started to appear that examples of anomalies seem to indicate that investors could potentially outsmart the market competition by identifying mispriced stocks following certain rules. Fundamental analysis of financial ratios are thereafter invented as a comparative tool in evaluating companies' business performance to identify those mispriced stocks.

However, the traditional method of fundamental analysis doesn't systematically relate financial ratios to stock return. Instead, its performance heavily relies on the experience of profession stock analytics. This study therefore aims to contribute to the field of stock investment by achieving two objectives 1) test the efficient market hypothesis and identify any financial ratios that have significant predictive power on future stock return. 2) build statistical models that can be used for creating stock investment portfolios that earn abnormal gains in stock markets if any useful relationship between financial ratios and stock return is identified.

Specifically, the study employs generalized linear model and random forest methods to investigate the relationship between 7 financial ratios and stock returns for 54 technology firms traded in the United States stock markets throughout 2010-2018. Two algorithms are applied in order to cross-check the reliability of the final results. Besides, to evaluate the practical value of the identified relationship, this study examines whether winning portfolios formed based on the identified relationship end up producing excess returns over losing portfolios and S&P 500 index using a hold-out testing set. Higher returns on winning portfolios would provide evidence for market anomalies that can be taken advantage of for stock investment and hence challenge efficient market hypothesis.

2 Methodology

2.1 Generalized Linear Regression

The term generalized linear models (GLMs) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors. GLMs are a broad class of models that include linear regression, ANOVA, Poisson regression, logistic regression, log-linear models, multinomial response, and etc. There are three components to any GLM:

- *Random Component* – refers to the probability distribution of the response variable Y; e.g. normal distribution for Y in the linear regression, or binomial distribution for Y in the binary logistic regression.
- *Systematic Component* - specifies the linear combination of explanatory variables X_1, X_2,…, X_k in the model, e.g., $\beta_0 + \beta_1 x_1 + \beta_2 x_2$ as in a linear regression, or in a logistic regression.
- *Link Function*, η or $g(\mu)$ - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g., $\eta = g(E(Y_i \mid X)) = E(Y_i \mid X)$ for linear regression, or $\eta = \text{logit}(\pi)$ for logistic regression.

GLM does not assume a linear relationship between the dependent variable and the independent variables, but it does assume a linear relationship between the transformed response in terms of the link function and the covariates; e.g., for binary logistic regression $logit(\pi) = \beta_0 + \beta X$. It requires the errors to be independent but not necessarily normally distributed. And it uses maximum likelihood estimation rather than ordinary least squares (OLS) to estimate the parameters, and thus relies on large-sample approximations.

2.2 Random Forests

Random forests are used as a way of building and improving tree-based models. Random forests build several decision trees on bootstrapped training samples. But each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those *m* predictors. This process is repeated to generate many different decision trees. When test data is fed to all of the decision trees, the algorithm maps input data to its predicted outcome by majority vote or by averaging the tree predictions.

Random forests provide an improvement over trees by decorrelating the trees. The rationale behind is that suppose that there is one very strong predictor in the data set, along with several other moderately strong predictors. Then in the collection of bagged trees, almost all of the trees will use this strong predictor in the top split. Consequently, all of the bagged trees will look quite similar. Hence the predictions from the bagged trees will be highly correlated. Random forests overcome this problem by forcing each split to consider only a subset of the predictors. Therefore, on average $(p - m) / p$ of the splits will not even consider the strong predictor, and so other predictors will have more of a chance. We can think of this process as decorrelating the trees, thereby making the average of the resulting trees less variable and hence more reliable.

The feature importance method for Random Forests is able to draw conclusions about what features contribute most to the decision making in the model. Feature importance is defined as its relative contribution to the decision making of the algorithm, and can be determined by different methods. Two common methods include (a). the mean decrease of accuracy in predictions on out-of-bag (OOB) samples when a variable is excluded from the tree, and (b). the decrease in training RSS due to splits for a predictor, averaged over all trees.

3 Data Summary

3.1 Data Description

The data used for this study is provided by SimFin's free database. The data contains seven separate datasets for the annual balance sheet, annual income statement, the annual statement of cash flow, daily stock price, industry information, and company information, sector information respectively. The final datasets comprise stock return and financial ratios calculated in annual terms based on the consideration that most companies publicize their financial statements annually.

Recognizing that the relationship between financial ratios and stock prices can be drastically different from markets to markets and industry to industry, this study will only focus on the technology sector of the United States, a new but growing sector with very high transparency in terms of sharing financial data. 54 technology companies are included, covering a period from 2008 to 2019. Selected firms vary in size and business modes and can be considered a good representation of technology stocks traded in the United States.

3.2 Inclusion and Exclusion Criteria

The companies included in the original SimFin datasets are filtered based on the following criteria:

(1) The companies must be listed in the United States stock exchange markets.

(2) The companies must be listed in the Technology sector.

(3) The companies must have been continuously traded from 2010 to 2019.

(4) The companies must have a positive book equity value.

(5) The companies must not have missing value in the columns needed for this study.

3.3 Variables of Study

Financial ratios used in this study were selected based on the consideration that they are widely used by professional investors and financial analysts in assessing six different aspects of a company's performance including market value, profitability, asset utilization efficiency, operating efficiency, capital structure, and liquidity. Since stock prices reflect investors' perceptions of companies' present value and future potentials, we believe the selected financial ratios should be able to potentially predict whether a company will grow in market value or not. Overall, 7 financial ratios and 2 stock return variables are included in the model. The descriptive statistics summary for all variables are shown in Table 3.1. For the generalized linear regression model, all input variables are in their logged form in order to fix the problem of skewness. For variables with negative values, log modulus transformation is used in instead: $L(x) = sign(x) * log(|x| + 1)$

Stock Return (R)

Two different stock return variables are calculated in this study. The first one is the raw return adjusted for dividend payments which is calculated according to the formula:

$$R_{a,i,t} = \frac{P_{i,t} - P_{i,t-1} + Div_{i,t}}{P_{i,t-1}}$$

$R_{a,i,t}$ means the percentage returns of stock i in year t; $P_{i,t}$ is the closing price of stock i at the end of year t; $P_{i,t-1}$ is the closing price of stock i at the start of year t, $Div_{i,t}$ is the dividend paid to shareholders of stock i in year t.

The second stock return variable is in logarithmic form. This variable is used to train the model because it has the merit of being more normally distributed than raw return. The following formula is used to calculate the variable:

$$R_{b,i,t} = log(\frac{P_{i,t}}{P_{i,t-1}})$$

Financial Ratios

7 financial ratios are used as explanatory variables in this study, below is their corresponding formula.

Book-to-Market Ratio (BM)	$BM_{i,t} = \frac{Total\ Equity_{i,t}}{P_{i,t} \times Outstanding\ Shares_{i,t}}$
Sales Per Share (SPS)	$SPS_{i,t} = \frac{Revenue_{i,t}}{Outstanding\ Shares_{i,t}}$
Return On Equity (ROE)	$ROE_{i,t} = \frac{Revenue_{i,t} - Income\ Tax\ Expense_{i,t} - Interest\ Expense_{i,t} - Depreciation_{i,t}}{Total\ Equity_{i,t}}$
Operating Profit Margin (OPM)	$OPM_{i,t} = \frac{Operating\ Income_{i,t}}{Revenue_{i,t}}$
Receivable Turnover (RT)	$RT_{i,t} = \frac{Revenue_{i,t} - Income\ Tax\ Expense_{i,t} - Interest\ Expense_{i,t} - Depreciation_{i,t}}{Account\ Receivable_{i,t-1}}$

Debt Ratio (DR)	$DR_{i,t} = \frac{Total\ Debt_{i,t}}{Total\ Assets_{i,t}}$
Current Ratio (CR)	$CR_{i,t} = \frac{Total\ Current\ Assets_{i,t}}{Total\ Current\ Liabilities_{i,t}}$

Variables	Notation	Mean	Median	SD	Min	Max
Raw Return	$\mathbf{R_a}$	0.4158	0.1262	5.7009	-0.5982	125.625
Log Return	$\mathbf{R_b}$	0.1189	0.1189	0.3463	-0.9118	4.8412
Book-to-Market Ratio	**BM**	0.6860	0.3056	3.5862	0.0097	43.5843
Sales Per Share	**SPS**	24.268	12.126	38.7233	1.730	335.543
Return On Equity	**ROE**	0.8578	0.4905	1.0012	0.0630	5.9629
Operating Profit Margin	**OPM**	0.1808	0.1818	0.1065	-0.3165	0.4934
Receivable Turnover	**RT**	8.227	6.897	5.3699	2.797	57.903
Debt Ratio	**DR**	0.4372	0.4300	0.1842	0.0699	0.9716
Current Ratio	**CR**	2.6038	2.1979	1.6354	0.6442	10.9466

Table 3.1: Descriptive statistics

4 Modeling & Analysis Results

4.1 Correlation Analysis

Before modeling, correlation analysis is done (1) to evaluate the strength of linear relationship between financial ratios and stock returns, and (2) to examine whether any financial ratios is highly correlated with other ratios and thus provide redundant information.

A correlation matrix for data used in generalized linear regression is constructed to achieve the purpose and shown below in Figure 4.1. The values of pearson correlation coefficients are shown within the boxes. The matrix shows that none of the explanatory variables seem to exhibit a strong correlation with stock returns. But hypothesis tests are still needed to test the statistical significance of the relationship between financial ratios and stock returns. For the problem of multicollinearity, none of the raitos has a correlation with another ratio of magnitude above 0.80. Therefore, the issue of multicollinearity among explanatory variables should not be a concern for the generalized linear regression model.

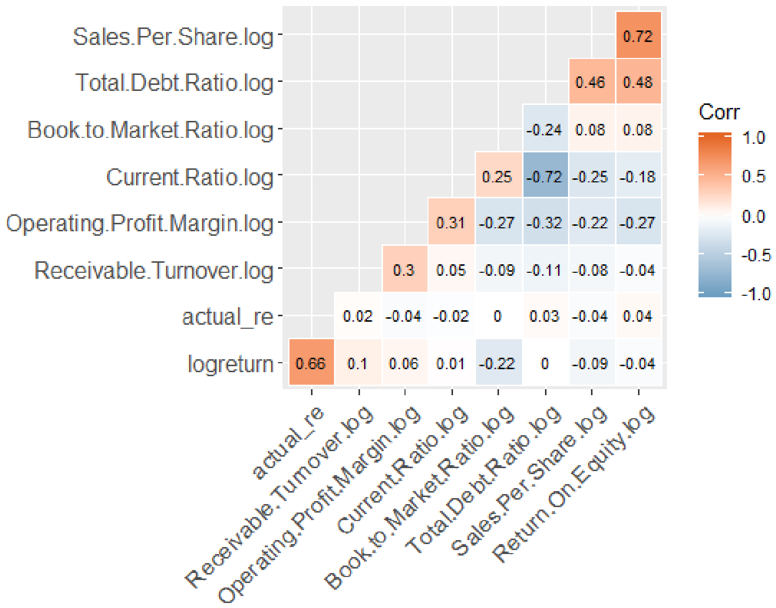


Figure 4.1

Note: financial ratios are in their log forms

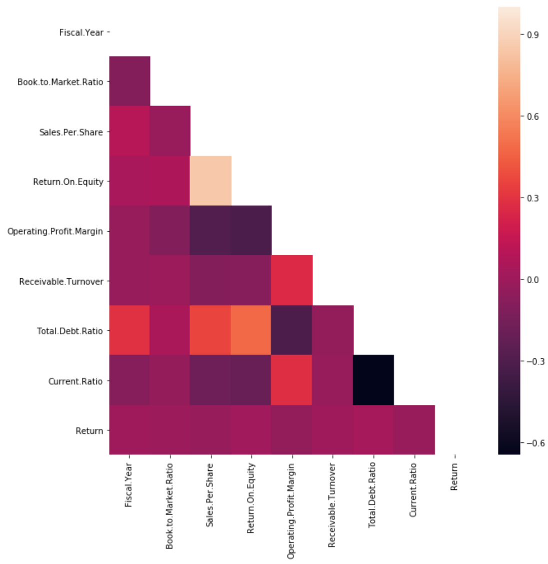


Figure 4.2

Note: financial ratios are in their original forms

A correlation matrix for data used in the random forest model is shown in Figure 4.2. According to the plot, return on equity has a positive correlation with Sales per Share and the correlation is greater than 0.8. Thus, for the random forest model, we remove the Return on Equity variable.

4.2 Generalized Linear Regression

We consider the following model for regression analysis:

$$R_{b,i,t} = \beta_0 + \beta_1 BM_{i,t} + \beta_2 SPS_{i,t} + \beta_3 ROE_{i,t} + \beta_4 OPM_{i,t} + \beta_5 RT_{i,t} + \beta_6 DR_{i,t} + \beta_7 CR_{i,t} + \varepsilon_{i,t}$$

$R_{b,i,t}$ = Log returns of firm i in year t

β_0 = Unsystematic constant component

β_n = Coefficients of the nth financial ratio, n=1,2,...,7

$\varepsilon_{i,t}$ = Unsystematic error from the predicted $R_{b,i,t}$

i=1, 2,…, 54

t=1, 2,…, 9

The theoretical framework presented above shows the response and explanatory variables of the model. The explanatory variables are book-to-market ratio, sales per share, return on equity, operating profit margin, receivable turnover, debt ratio, and current ratio. The response variable is the log stock return. It is important to note that the financial ratios used for regression analysis are in their log-transformed form because logarithm helps normalize skewed distribution, which is a common problem for ratio data, and improve the performance of linear regression.

The results of multiple linear regression based on all data from 2010 to 2018 are given in table 4.1. The finding reveals that only the book-to-market ratio is statistically significant at the 1% significance level. The book-to-market ratio is negatively associated with log stock return, suggesting that stocks with relatively high market value outperform stocks with a low market value in the stock market. The finding that book-to-market ratio can be used to pick out high return stock from low return ones sheds light on the book-to-market effect, one of the oldest effects that have been investigated in financial markets. But the relationship shown in this study is in reverse to the classic book-to-market effect. The book-to-market effect states that portfolios with high book-to-market ratio stocks usually exhibit a higher average return than those with low book-to-market ratio because they are usually considered as being undervalued and thus have higher potential. The regression analysis here shows a different relationship but still provides evidence for the presence of market anomaly and thus further challenges the efficient market hypothesis.

However, in terms of the predictive power measured by the adjusted R square , the overall adjusted R square of this multiple linear regression is merely 4.83%, which means only 4.83% of the variation in log stock return is explained by the seven financial ratios. After removing all other ratios and running simple linear regression of log stock return on book-to-market ratio, the adjusted R square was reduced to 4.41%. The reduction is very small, suggesting that book-to-market ratio is dominant in terms of explaining variations in log stock returns among all ratios while other ratios contribute very little to the prediction.

Predictor	**Estimated Coefficient**	**t Satistics**	**P-value**
BM	-0.09637 (0.02121)	-4.543	7.02e-06***
SPS	-0.03606 (0.02546)	-1.416	0.157
ROE	0.01167 (0.02735)	0.427	0.670
OPM	-0.22031 (0.20652)	-1.067	0.287
RT	0.06695 (0.03765)	1.778	0.076*
TDR	0.04220 (0.13281)	0.318	0.751
CR	0.05002 (0.04323)	1.157	0.248
Adj.R^2: 0.04829	F-statistic: 4.516 on 7 and 478 DF	p-value: 6.859e-05	
Note: 1. ***, ** and * indicate significance at 1%, 5%, and 10% respectively. 2. Standard error of estimated coefficient is reported in the parenthesis.			

Table 4.1

From 2010 to 2018, the stock market of the United States has undergone several bull and bear cycles. Taking that into consideration, it is reasonable to suspect that the underlying relationship between stock returns and financial ratios may vary across years. So the same regression was run again on data of each year respectively in order to examine whether the book-to-market ratio has consistently been a significant predictor of stock return under different economic conditions. The corresponding results are shown in Table 4.2. The results disclose that in 2012,2013, 2014, and 2016, the book-to-market ratio is statistically insignificant in predicting stock return. Also, in 2012, 2014, 2016, and 2018, other financial ratios including sales per share, return on equity, and current ratio also showed strong predictive power on stock return, even stronger than the book-to-market ratio. These findings suggest that although investing in stocks with high book-to-market ratios could lead to beating the market, it doesn't guarantee continuous success which partially supports the efficient market hypothesis.

Predictor	2010	2011	2012	2013	2014	2015	2016	2017	2018
BM	-0.117 ***	-0.085*	-0.045	-0.012	0.192	-0.204 **	-0.111	-0.109*	-0.166 ***
SPS	-0.015	0.002	0.087*	-0.044	-0.414*	0.012	-0.007	-0.059	0.121*
ROE	-0.058	-0.014	-0.067	0.079	0.350 *	-0.027	0.058	0.006	-0.166 **
OPM	-0.630	-0.039	0.395	-0.143	0.086	0.109	-0.652	-0.029	-0.287
RT	0.059	0.127	-0.036	0.083	0.157	-0.015	0.103	0.067	-0.017
TDR	-0.181	0.292	0.031	-0.147	0.220	-0.095	-0.040	-0.150	0.035
CR	-0.028	0.074	-0.005	0.030	-0.091	-0.100	0.241*	0.023	-0.021
Note: ***, ** and * indicate significance at 1%, 5%, and 10% respectively.									

Table 4.2

According to the results of the regression analysis shown above, although book-to-market is statistically significant in predicting stock returns in most years, no financial ratios seem to have a consistent and robust predictive power on stock returns. The very small adjusted R square also indicates that most variation in stock returns cannot be explained by financial ratios. Therefore, although the regression analysis does provide evidence for a significant relationship between book-to-market ratio and stock return and suggests chances of outperforming the market in stock investment, in general the regression analysis fails to reject the efficient market hypothesis.

4.3 Random Forests

4.3.1 Data processing

We build the model by using the same data set as generalized linear regression does, which includes no missing values in the selected years and variables. However, we do not conduct the log transformation since a random forest is invariant to monotonic transformation. As for the predictor variables, we use the 6 variables without multicollinearity issues. These six variables are: book to market ratio, sales per share, operating profit margin, receivable turnover, debt ratio and current ratio. For the validation purpose, we split the data into training (2010-2015) and test data(2016-2018). The response variable is the total stock return.

4.3.2 Feature importance

The feature importance for the random forest model is as follows:

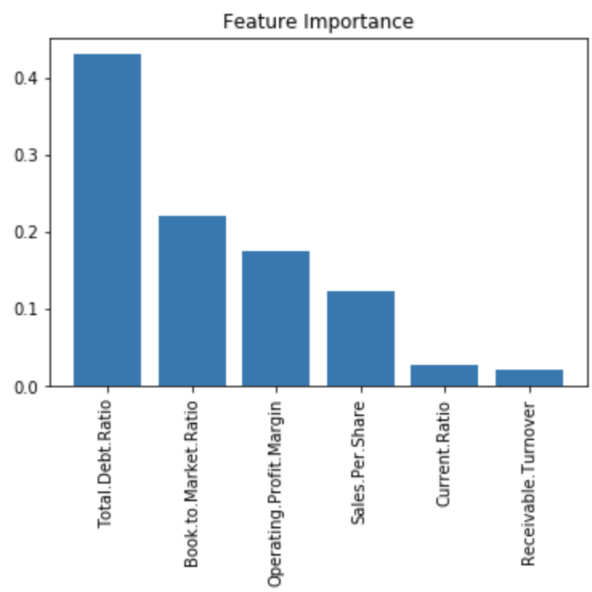


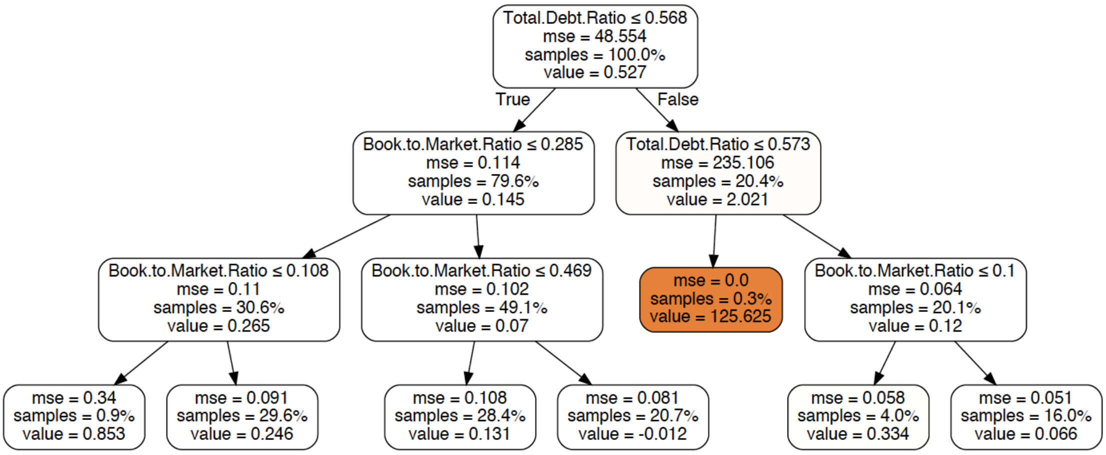
Figure 4.2

According to figure 4.2, total debt ratio has the highest feature importance (0.43) and book to market ratio has the second largest feature importance (0.22). This coincides with the result in the decision tree, which is introduced in the next session.

It is also worthy to note that while total debt ratio is the most important feature, current ratio is the second least important feature. This means that a company's long-term leverage is more important than short-term liquidity in determining stock return.

4.3.3 Decision tree

Decision trees are easy to compute and explain why a particular variable is having higher importance. The tree can be visualized and hence, for non-technical users, it is easier to explain model implementation. A decision tree is built on an entire dataset, using all the features / variables of interest, whereas a random forest randomly selects observations and specific features to build multiple decision trees and then average the results. I set maximum depth of decision tree to be 3 and the result is as follows:



According to the tree, there are three major findings:

1. Total debt ratio is the most important feature and book to market ratio is the second most important feature in determining the company's total stock return.
2. Companies with low total debt ratio and high book to market ratio tend to have negative total stock return, meaning that undervalued companies with relatively low debt are more likely to have negative stock return. This coincides with the result of generalized linear regression.
3. 20.4% companies' total debt ratio is higher than 0.568, and all of those companies have high total stock returns.The average total debt ratio for the technology industry worldwide is 0.252 and Facebook's total debt ratio is 0.093. This means tech companies with higher than average debt (leverage) and high risk are more likely to

have high stock returns. This makes sense since high debt also means the company is financing increased operations through borrowing, which means the company is growing.

4.3.3 Validation

For the random forests algorithm, this study employs three validation strategies: MSE, correctness of price movement, and comparison between good and bad portfolios. The random forest model has a MSE of 2.23 which is fairly low; The decision tree has an even lower MSE of 0.12. One possible explanation is as follows: total debt ratio is truly an important feature. While using random forest, total debt ratio is not guaranteed to be included in each tree, thus making the final prediction less reliable. However, in the decision tree, total debt ratio is used as the root node and constantly appeared as internal nodes. Another possible reason is that decision trees tend to suffer from variance error, meaning that a slight change in the training data set might exert a huge difference to the final output. It is true that the decision tree worked perfectly in the existing dataset, but whether this is true for data in further years remains unknown.

As for price movement, we check what percentage of the predicted returns move in the same direction as the actual returns. The accuracy of prediction would be 50% with random guessing. Using the random forest model, the accuracy of prediction increases to 67.28%; However, our model fails the third validation method. A detailed explanation is provided in 4.4.

4.4 Investment Strategies

In order to assess the practical value of using financial ratios to predict stock returns, we created 6 stock investment portfolios, one winning portfolio and one losing portfolio for each period of 2016, 2017, and 2018, respectively. Each year's log stock return is predicted using data from previous years. Portfolios are then constructed based on the predicted log stock return. 54 stocks are ranked according to their estimated returns and 10 stocks with the highest estimated return form the winning portfolio while 10 stocks with the lowest returns form the losing portfolio. The actual returns for the stocks in each portfolio are then averaged to calculate the portfolio returns. The performance of portfolios constructed with generalized linear regression is presented in Table 4.3. The performance of portfolios constructed with random forests is presented in Table 4.4.

	2016	2017	2018
Winning Portfolio	52.78%	62.06%	11.20%
Losing Portfolio	30.29%	5.52%	-20.19%
Excess Return	22.49%	56.54%	31.39%
NASDAQ 100 Technology Sector return	26.87%	35.64%	-6.57%

Table 4.3 - Resulted portfolios from generalized linear regression

	2016	2017	2018
Winning Portfolio	15.75%	28.08%	-19.09%
Losing Portfolio	32.57%	24.01%	-17.20%
Excess Return	-16.82%	4.07%	-1.89%

Table 4.4 - Resulted portfolios from random forests

The results displayed via table 4.3 show that winning portfolios outperform the losing portfolios for all three years. In 2017, the excess return of the winning portfolio over the losing portfolio is the greatest which reaches 56.54%. Even in 2016, the year in which the excess return is the smallest, the winning portfolio still outperforms the losing portfolio by 22.49%. The return of the winning portfolio is significantly higher than the return of the NASDAQ 100 Technology Sector in all three years, with an average increase of 23.37%. These results suggest that the investment strategy guided by financial ratios under regression analysis is able to produce higher than average returns in the short term even though such chances of profitability can be easily shackled in the long term.

The results displayed via table 4.4 shows that the winning portfolios only outperform the losing portfolio in 2017 with an excessive return as small as 4.07%. In 2016 and 2018, the winning portfolio has a lower return than the losing portfolio, meaning that the investment strategy guided by random forests is unreliable.

5 Conclusion & Recommendation

5.1 Conclusion

This paper explored the relationship between financial ratios and stock returns using a data set of 54 technology companies over the period from 2010 to 2018. Our model fails to discover a consistent relationship between financial ratios and stock returns over the 9 year period of interest. This makes sense since from 2010 to 2018, the market has gone through several bull and bear cycles. However, we do prove that in specific years including 2010, 2011, 2015, 2017, and 2018, the book-to-market ratio is statistically significant in predicting stock returns and holds a negative association. And in 2012, 2014, 2016, and 2018, financial ratios other than book-to-market ratio, including sales per share, return on equity, and current ratio, also exhibit strong predictive power on stock return.

Our models successfully disprove the weak efficient marketing hypothesis from two angles. First, by intelligently forming a winning portfolio with 10 stocks using generalized linear regression, investors are expected to get a return on average 26.67% higher than NASDAQ 100 Technology Sector index. Second, the random forest model predicts the direction of market price movement with an accuracy of 67.28%, which is significantly higher than random guessing.

It is hard to compare the performance of the two models since they use different validation methods. GLM assumes a linear relationship of the data, so it uses R^2, whereas random forests and decision trees use MSE since it does not assume a linear relationship. Additionally, the two models are used in different scenarios. The GLM model is valuable for investors who are seeking suggestions on investment portfolios that guarantee higher annual returns than NASDAQ-100 Technology Sector Index. The random forests model is suitable for investors who already have an investment strategy in their mind and just want to check if their strategy will earn them positive returns.

5.2 Limitation and Future Directions

One constraint of our analysis is that we only include historical numerical data in the model to predict the stock price,which only disproves the weak efficient marketing hypothesis. However in reality, the stock price might also be affected by numerous qualitative factors including macroeconomic factors and social media presence. We have two recommendations for the future research on stock price prediction: 1) use text analysis for social media interpretation and explore the relationship between social media presence and stock price change. 2)create indices that quantify the macroeconomic variables including confidence towards the management board and change in trade policy.

One drawback of using generalized linear regression is that it does not account for possible nonlinear relationship between financial ratios and stock returns. Future research can try relaxing the linearity assumption while maintaining the positive aspects of linear regression. Some nonlinear models to consider include polynomial regression, regression splines, and smoothing splines. If time allows, we would also like to improve our linear regression model by trying shrinkage methods like ridge regression and Lasso, which can potentially reduce the variance so that our model will be more robust against stock market volatility. Given the time constraint, we only focus on companies in the technology sector. Future study can be extended to explore companies in other sectors and see if the models still work.

6. Reference

1. https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp
2. https://pdfs.semanticscholar.org/0c0f/bdfb8b82795fb3e7623bce087d377921c2c9.pdf
3. https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/4906/Predictability_of_stock_returns-Greek_stock_market-2010.pdf?sequence=2
4. http://www.ijtef.org/papers/137-S00072.pdf
5. https://www.investopedia.com/terms/o/operatingmargin.asp
6. https://www.investopedia.com/terms/r/receivableturnoverratio.asp
7. https://online.stat.psu.edu/stat504/node/216/