Key Financial Indicators: Unlocking Stock Performance Insights

Radjamin Hukom

rhukom@bellarmine.edu

4/29/2025

Executive Summary

This project investigates the predictive power of financial ratios in forecasting stock price changes over a 5-year period (2019–2024) using a dataset of 503 S&P 500 company stocks. Initially compiled through Finsheet.io and finalized using Bloomberg's API, the dataset includes over 62 financial variables per firm, alongside the percentage change in stock price as the response variable. Extreme values, skewed distributions, and high variance inflation factors (VIF) highlighted the dataset's inherent challenges, particularly multicollinearity and heteroscedasticity.

Several regression models were tested, including Multiple Linear Regression (MLR), Ridge, Lasso, Elastic Net, Support Vector Regression, Random Forest, Decision Tree, K-Nearest Neighbors, Weighted Least Squares (WLS), and Logistic Regression. Early results showed only moderate predictive power, except for 2019, which consistently outperformed other years across multiple models. Further analysis revealed significant model violations including high VIF scores and heteroscedastic residuals.

To improve model robustness, variable transformations were applied: related features were averaged into broader metrics—Returns, Liquidity, and Efficiency—and highly collinear variables like Operating and Net Profit Margins were dropped. Additional diagnostic testing confirmed persistent multicollinearity and heteroscedasticity even after these adjustments. Expanding the dataset by adding 1,000 more observations for 2019 surprisingly worsened model conditions due to increased variance.

Ultimately, the model achieved robustness only after removing beta and EPS (2019)—two highly influential but problematic predictors. This final version exhibited no multicollinearity (VIF < 1.2 for all variables), no heteroscedasticity (Breusch-Pagan p-value ≈ 0.95), and no autocorrelation (Durbin-Watson ≈ 2.00), while still maintaining an exceptionally high adjusted R^2 of 0.9999999999991485 using Ridge Regression. These results suggest that while beta and EPS are powerful predictors, their inclusion may compromise model stability in aggregate-level prediction. The findings support the use of composite financial metrics and robust diagnostics in financial forecasting.

Introduction

This project explores the use of predictive analytics in modeling stock performance using financial ratios, with the goal of identifying which factors most significantly influence stock price changes over time. Specifically, the objective was to identify key financial factors in predicting the performance of stocks over a 5-year period and to construct a robust, reliable predictive model capable of overcoming common econometric issues such as multicollinearity, heteroscedasticity, and autocorrelation.

The dataset includes 503 observations of S&P 500 stocks, with detailed annual financial data from 2019 to 2023, including variables such as Earnings Per Share (EPS), Return on Assets (ROA), liquidity ratios, and turnover metrics. The project uses a mix of traditional and modern regression models and applies advanced preprocessing steps—feature engineering, transformations, and regularization—to enhance model reliability. Through careful refinement and diagnostics, the project culminates in a highly robust Ridge Regression model with strong predictive capacity and minimal statistical violations.

Project Summary

About the Dataset

	Beta	EPS (2019)	ROE (2019)	ROA (2019)	Gross Margin (2019)	Operating Margin (2019)	Net Profit Margin (2019)	Cash Ratio (2019)	Current Ratio (2019)	Quick Ratio (2019)	...	Gross Margin (2023)	Operating Margin (2023)	Net Profit Margin (2023)	Cash Ratio (2023)	Current Ratio (2023)	Quick Ratio (2023)	A-Turnover (2023)	I-Turnover (2023)	R-Turnover (2023)	% change, Price
count	503.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	...	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000	502.000000
mean	0.797630	118.382686	-0.153562	0.067533	0.424493	0.143821	0.112534	0.571659	1.511457	1.141881	...	35.789144	16.426062	13.066706	0.527319	1.404682	0.892780	0.680264	8.436862	11.903914	0.776320
std	0.584612	2523.275699	5.609127	0.079839	0.266557	0.442281	0.411438	1.262977	1.434420	1.157206	...	26.779962	17.497430	17.264114	1.114254	1.480181	1.288385	0.613431	23.766885	28.616814	1.634994
min	-0.393969	-9.604000	-104.043480	-0.363637	0.000000	-9.116667	-8.566667	0.000000	0.000000	0.000000	...	-29.881300	-93.821700	-96.559400	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.841800
25%	0.377332	1.739999	0.076486	0.024370	0.232857	0.081935	0.065771	0.083174	0.768950	0.503162	...	12.362475	8.923800	5.986450	0.055775	0.749650	0.269900	0.301950	0.000000	4.633000	0.050575
50%	0.728988	3.685285	0.143494	0.055565	0.424897	0.149117	0.117623	0.253207	1.212050	0.901789	...	35.311100	16.220200	12.056300	0.228000	1.136750	0.673150	0.540800	2.832150	6.583650	0.413550
75%	1.084444	6.253253	0.254951	0.102909	0.639748	0.220374	0.190601	0.588211	1.827400	1.415405	...	57.851425	24.378550	19.393800	0.578875	1.668925	1.056275	0.824300	5.818025	10.561750	0.977675
max	4.982928	56539.582000	20.230770	0.532833	1.000000	0.876621	0.649807	19.666666	15.330200	13.173799	...	93.116700	101.644300	101.236000	16.337400	19.430600	19.243000	4.744400	268.852500	381.935000	21.838400

8 rows × 62 columns

Ratio	Meaning	E(sign)
Beta	Indicates the stock's risk relative to the market	-
EPS	Shows profitability per share	+
ROE	Measures how efficiently equity generates profit	+
ROA	Reflects how effectively assets generate profit	+
Gross Margin	Reveals profitability after production costs	+
Operating Margin	Indicates how well operations generate profit before interest and taxes.	+
Net Profit Margin	Shows overall profitability after all expenses	+
Cash Ratio	Assesses short-term liquidity	+
Current Ratio	Measures short-term financial health	+
Quick Ratio	Evaluates liquidity without relying on inventory	+
Asset Turnover	Shows how efficiently assets generate revenue	+
Inventory Turnover	Reflects how quickly inventory is sold	+
Receivables Turnover	Measures how fast credit sales are collected	+

The dataset was found first through the usage of finsheet.io, a free tool with pay-to-use add-ins to export financial data to excel and google sheets. However, it became clear that it is not powerful enough to handle the volume of requests needed to build this dataset. The rest of the data was found through Bloomberg, a well-respected software used by financial specialists all around the world for years that the university has paid for its services and API (Application Programming Interface) with the help of a donor. The table above gives an explanation of the meaning of each financial metric recorded in the dataset and its expected sign based on theory of its correlation with the response variable, the percentage change in price of the ticker symbol's stock over the period of 2019 to 2024.

The dataset consists of 503 observations with Ticker + Beta + 12 financial ratios * 5 years = 62 financial variables, providing a comprehensive overview of firm performance across multiple years. Including the response variable, % Price Change, we have 63 columns and 503 rows of stocks from S&P500. 503 ≠ 500 as some companies in the S&P500 may have more than one ticker. The mean values of key financial ratios indicate the overall industry trends, such as an average EPS of 118.38 and an ROA of 0.0675, suggesting that firms generally maintain positive, albeit modest, returns on assets. However, the standard deviations for these variables, especially EPS (2523.28) and ROE (5.61), highlight substantial variability among firms, pointing to a mix of highly profitable companies and those experiencing significant financial struggles. The presence of extreme values in these variables suggests that some firms generate exceptionally high earnings, while others report heavy losses, contributing to the dataset's high skewness.

Examining the quartiles further supports this variability. The median EPS is only 3.69, and the 75th percentile is 6.25, yet the maximum EPS reaches an astounding 56,539.58, indicating that a few firms are driving the high mean value. Similar patterns emerge for liquidity and turnover ratios. For instance, the cash ratio ranges from 0 to 19.67, with a median of 0.25, reflecting that while most firms maintain conservative liquidity levels, a few hold unusually high cash reserves. Additionally, turnover ratios such as Inventory Turnover (I-Turnover) and Receivables Turnover (R-Turnover) exhibit large spreads, with maximum values reaching 268.85 and 381.94, respectively, suggesting that certain firms operate with highly efficient turnover cycles while others struggle with slow-moving assets.

The dataset's extreme minimum and maximum values further highlight the presence of outliers. Some firms report highly negative financial metrics, such as an ROE as low as -104.04 and a net profit margin reaching -96.56, indicating significant financial distress. On the other hand, certain firms exhibit extraordinarily high performance,

with operating margins peaking at 101.64 and net profit margins at 101.24, which may be anomalies or cases of exceptionally profitable entities. The presence of such disparities suggests that data transformation techniques, such as normalization or winsorization, may be necessary to ensure a more balanced dataset suitable for predictive modeling.

On the Data Values

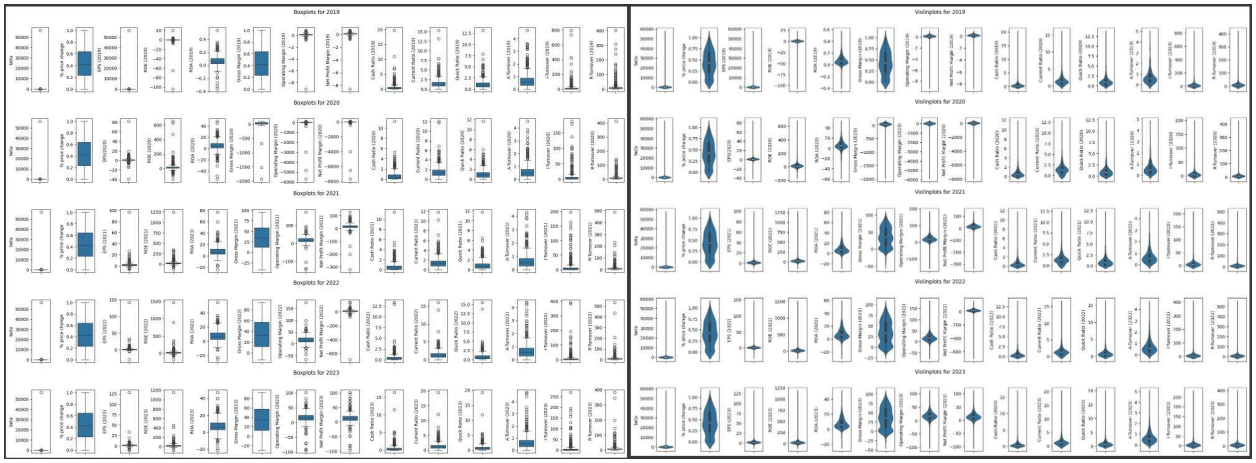
	Year	Column	Column Type	Outlier Count	Outlier Percentage
0	2019	EPS (2019)	ratio	38	7.54%
1	2019	ROE (2019)	ratio	67	13.29%
2	2019	ROA (2019)	ratio	29	5.75%
3	2019	Gross Margin (2019)	ratio	0	0.00%
4	2019	Operating Margin (2019)	ratio	29	5.75%
5	2019	Net Profit Margin (2019)	ratio	28	5.56%
6	2019	Cash Ratio (2019)	ratio	38	7.54%
7	2019	Current Ratio (2019)	ratio	37	7.34%
8	2019	Quick Ratio (2019)	ratio	35	6.94%
9	2019	A-Turnover (2019)	ratio	37	7.34%
10	2019	I-Turnover (2019)	ratio	60	11.90%
11	2019	R-Turnover (2019)	ratio	57	11.31%
12	2020	EPS(2020)	ratio	50	9.92%
13	2020	ROE (2020)	ratio	56	11.11%
14	2020	ROA (2020)	ratio	44	8.73%
15	2020	Gross Margin (2020)	ratio	3	0.60%
16	2020	Operating Margin (2020)	ratio	37	7.34%
17	2020	Net Profit Margin (2020)	ratio	47	9.33%
18	2020	Cash Ratio (2020)	ratio	27	5.36%
19	2020	Current Ratio (2020)	ratio	29	5.75%
20	2020	Quick Ratio (2020)	ratio	29	5.75%
21	2020	A-Turnover (2020)	ratio	29	5.75%
22	2020	I-Turnover (2020)	ratio	46	9.13%
23	2020	R-Turnover (2020)	ratio	48	9.52%
24	2021	EPS (2021)	ratio	30	5.95%
25	2021	ROE (2021)	ratio	43	8.53%
26	2021	ROA (2021)	ratio	24	4.76%
27	2021	Gross Margin (2021)	ratio	0	0.00%
28	2021	Operating Margin (2021)	ratio	22	4.37%
29	2021	Net Profit Margin (2021)	ratio	29	5.75%
30	2021	Cash Ratio (2021)	ratio	36	7.14%
31	2021	Current Ratio (2021)	ratio	28	5.56%
32	2021	Quick Ratio (2021)	ratio	27	5.36%
33	2021	A-Turnover (2021)	ratio	31	6.15%
34	2021	I-Turnover (2021)	ratio	47	9.33%
35	2021	R-Turnover (2021)	ratio	48	9.52%
36	2022	EPS (2022)	ratio	36	7.14%
37	2022	ROE (2022)	ratio	39	7.74%
38	2022	ROA (2022)	ratio	25	4.96%
39	2022	Gross Margin (2022)	ratio	0	0.00%
40	2022	Operating Margin (2022)	ratio	22	4.37%
41	2022	Net Profit Margin (2022)	ratio	30	5.95%
42	2022	Cash Ratio (2022)	ratio	44	8.73%
43	2022	Current Ratio (2022)	ratio	32	6.35%
44	2022	Quick Ratio (2022)	ratio	25	4.96%
45	2022	A-Turnover (2022)	ratio	33	6.55%
46	2022	I-Turnover (2022)	ratio	54	10.71%
47	2022	R-Turnover (2022)	ratio	43	8.53%
48	2023	EPS (2023)	ratio	46	9.13%
49	2023	ROE (2023)	ratio	35	6.94%
50	2023	ROA (2023)	ratio	26	5.16%
51	2023	Gross Margin (2023)	ratio	0	0.00%
52	2023	Operating Margin (2023)	ratio	23	4.56%
53	2023	Net Profit Margin (2023)	ratio	35	6.94%
54	2023	Cash Ratio (2023)	ratio	41	8.13%
55	2023	Current Ratio (2023)	ratio	41	8.13%
56	2023	Quick Ratio (2023)	ratio	30	5.95%
57	2023	A-Turnover (2023)	ratio	34	6.75%
58	2023	I-Turnover (2023)	ratio	57	11.31%
59	2023	R-Turnover (2023)	ratio	46	9.13%

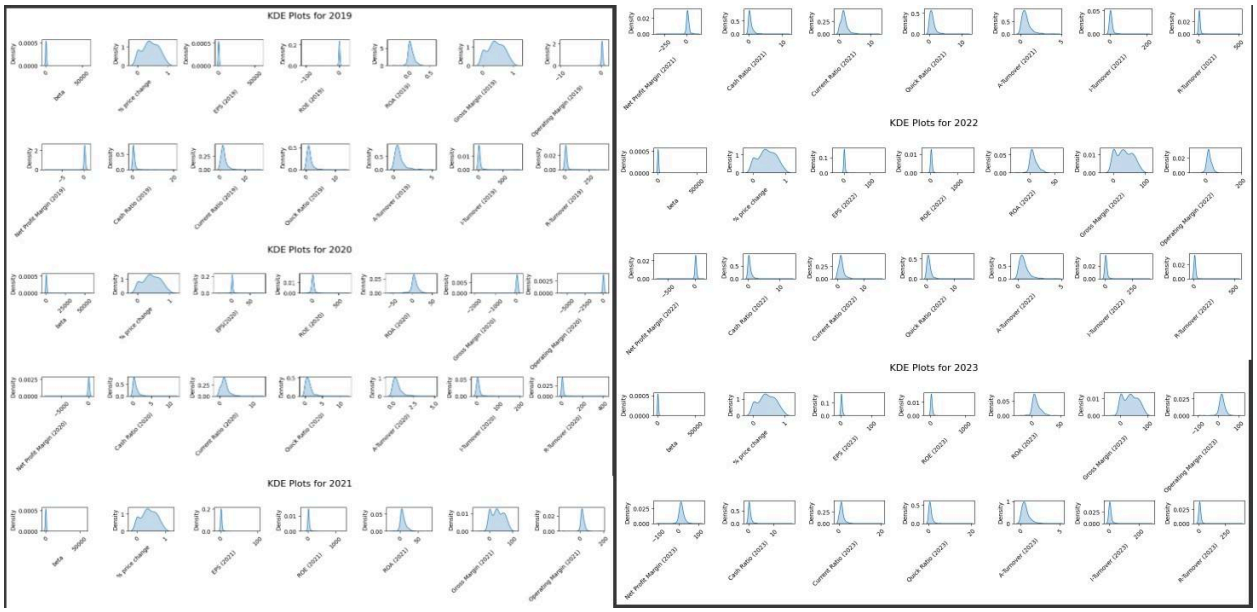
	Column Name	Column Type	Count of Outliers	% of Outliers
0	Ticker	nominal	0	0.00%
1	beta	ratio	38	7.55%
2	% price change	ratio	0	0.00%

The dataset contains a significant number of outliers, particularly in financial ratios such as EPS, ROE, ROA, and turnover metrics, with outlier percentages exceeding 10% in some cases. Notably, gross margin remains relatively stable across all years with minimal outliers. High outlier counts in key profitability and efficiency ratios suggest substantial variability in company performance, which could pose challenges for using Multiple Linear Regression (MLR). Outliers can distort coefficient estimates, reduce model accuracy, and increase heteroscedasticity. Despite the presence of extreme values, it has been determined that outliers should be retained due to the inherent characteristics of financial data. Unlike in other domains where outliers may result from data entry errors or measurement inconsistencies, financial data often exhibits genuine extreme values driven by market

conditions, firm-specific events, or economic cycles. For example, companies experiencing rapid growth, financial distress, or industry disruptions may naturally produce extreme EPS, ROE, or margin figures. Removing these outliers could eliminate crucial insights about firms that significantly outperform or underperform their peers, leading to a misrepresentation of real-world financial dynamics.

Distribution Analysis

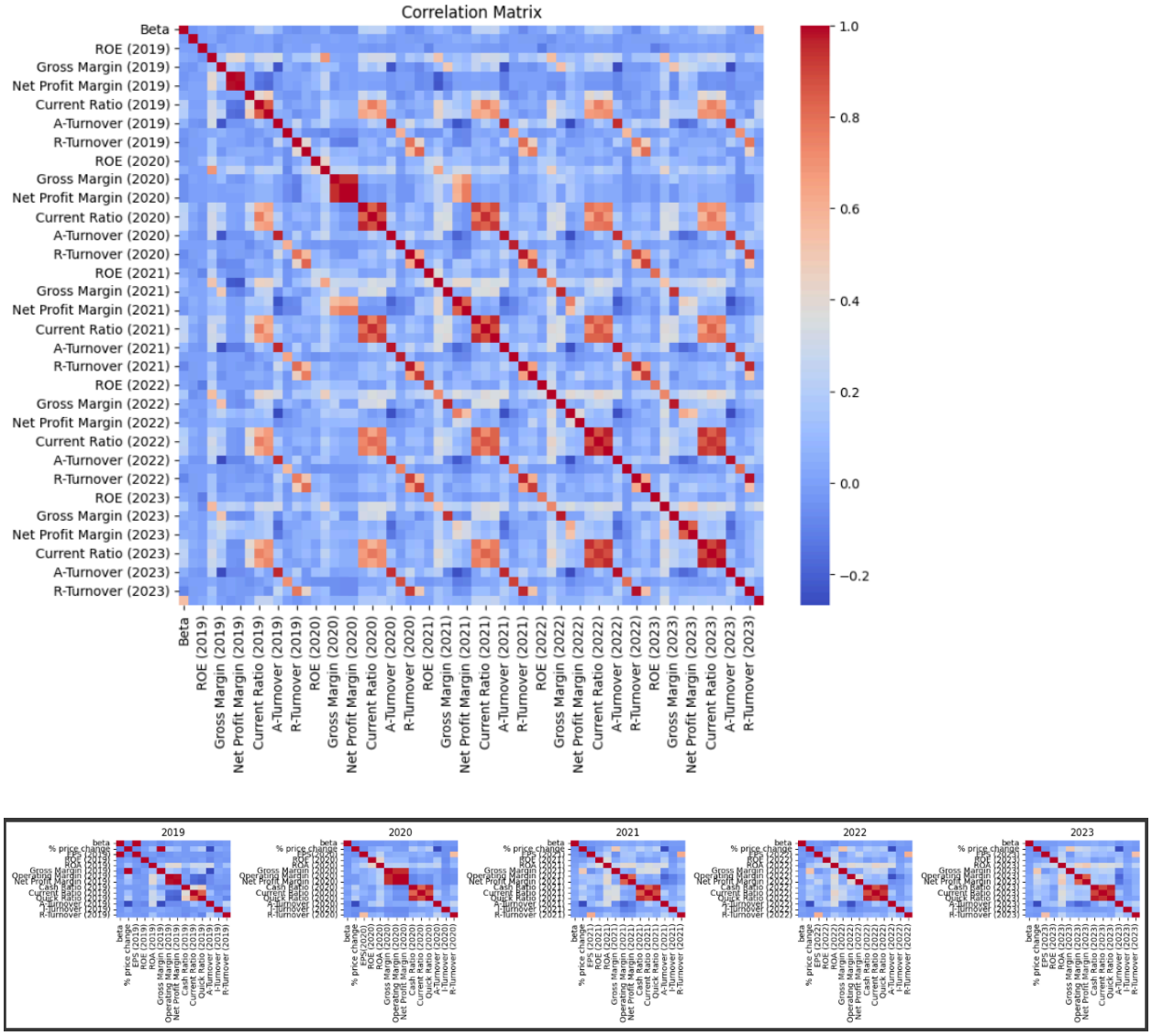




The dataset exhibits significant skewness across multiple financial ratios, which can negatively impact the assumptions of normality required for Multiple Linear Regression (MLR). Notably, EPS and ROE display extreme skewness across all years, with values exceeding ±10 in several instances, suggesting the presence of outliers that could distort regression results. Additionally, net profit and operating margins show substantial negative skewness, indicating that most firms have relatively small or negative margins, while a few extreme positive values pull the

mean downward. Liquidity ratios, such as the Cash Ratio, Current Ratio, and Quick Ratio, also exhibit considerable positive skewness, meaning some firms have abnormally high liquidity compared to the rest of the sample. Similarly, turnover ratios, including Inventory Turnover (I-Turnover) and Receivables Turnover (R-Turnover), tend to have high positive skewness, reflecting a few firms with exceptionally high turnover levels.

Correlation Analysis



The correlation analysis from 2019 to 2023 highlights Gross Margin as the most consistently significant factor, showing strong positive correlations with percentage price changes across all years. Other financial metrics, such as Operating Margin, Net Profit Margin, Quick Ratio, and Cash Ratio, exhibit occasional significance but with weaker and less consistent correlations. Interestingly, EPS, ROE, and ROA generally have low or negligible

correlations, suggesting they may not be strong predictors of price changes in this dataset. The presence of multicollinearity, particularly with highly correlated variables like Gross Margin, could affect the reliability of multiple linear regression (MLR) by inflating variance and reducing the precision of coefficient estimates. This suggests that while MLR can still provide insights, alternative techniques such as principal component regression (PCR) or ridge regression may help mitigate multicollinearity and improve model robustness.

Preliminary Attempts at Predictive Analytics

Data and Preprocessing

- The dataset included company tickers, beta values, and various financial ratios
- Year-specific data was isolated into five datasets: df_2019 to df_2023.
- The % price change column served as the main response variable, and a binary target +/- price change was also created for classification models.
- Column names were standardized and grouped into financial themes (e.g., Return, Liquidity, Efficiency) to address multicollinearity.

Machine Learning Models

1. **Multiple Linear Regression (MLR)**: Establishes baseline relationships between variables.
2. **Ridge & Lasso Regression**: Enhances model stability by regularizing coefficients.
3. **Elastic Net Regression (ENR)**: Balances Ridge and Lasso benefits for correlated predictors.
4. **Support Vector Regression (SVR)**: Captures complex non-linear stock trends.
5. **Random Forest Regression (RFR)**: Uses ensemble learning to improve accuracy.
6. **Decision Tree Regression (DTR)**: Simple but prone to overfitting.
7. **K-Nearest Neighbors Regression (KNNR)**: Predicts based on similar stock behaviors.
8. **Weighted Least Squares (WLS) Regression**: Handles heteroscedasticity in financial data.
9. **Logit**: Predicts positive or negative changes over a 5-year period.

Model	2019	2020	2021	2022	2023
Multiple Linear Regression (MLR)	1.000	~0.25	~0.33	~0.37	~0.30

Ridge Regression	0.9990	0.2355	~0.31	0.3586	~0.32
Lasso Regression	< 0	< 0	< 0	< 0	< 0
Elastic Net Regression	< 0	~0.24	~0.29	~0.29	~0.28
Support Vector Regression (SVR)	< 0	< 0	< 0	< 0	0.4743
Random Forest Regression	0.9992	~0.22	~0.28	~0.31	0.3349
Decision Tree Regression	0.9989	Low/Negative	< 0	< 0	< 0
K-Nearest Neighbors (KNN)	< 0	~0.21	~0.12	~0.18	0.2124
Weighted Least Squares (WLS)	1	~0.22	~0.29	~0.30	0.3443
Logistic Regression (Logit)	.9663	0.3143	0.3407	0.3408	0.325

Seeing these results, it confirmed my previous fears of the dataset having inbuilt problems with robustness, whether it is multicollinearity, heteroscedasticity, or autocorrelation. Of course, further diagnostic was done to further confirm which of these three things did my results violate.

Diagnostic Testing

- **Multicollinearity:** VIF > 40 for some predictors (e.g., Operating Margin), indicating redundancy. Variables were dropped or combined.
- **Heteroscedasticity:** Present in most years, corrected with WLS.
- **Autocorrelation:** Durbin-Watson tests showed no significant autocorrelation.
- **Ramsey RESET Tests:** No evidence of model misspecification for MLR; however, Logit results in 2019 were misspecified.

First Round of Refinement (Manipulation of Variables)

Seeing the year 2019 consistently outperforming the others on the models that understood the data (with nonnegative R-squared), in the interest of time, I focused my efforts on making df_2019 robust.

Model Implementation

- **Multiple Linear Regression (MLR):** Used ordinary least squares with adjusted R^2 values calculated per year.
- **Regularization Techniques:**
    - **Ridge Regression:** Applied to mitigate overfitting.
    - **Lasso Regression:** Employed for feature selection.
    - **Elastic Net:** Balanced Ridge and Lasso effects.

Multicollinearity Handling

- **Variance Inflation Factor (VIF):** Identified highly correlated predictors.
- **Feature Engineering:** Grouped variables into combined metrics such as Returns (equal to mean of ROA, ROE), Liquidity (mean of Current, Quick, Cash Ratios), and Efficiency (mean of Turnover ratios).
- **Removal of Redundant Predictors:** Operating Margin and Net Profit Margin were dropped due to excessive collinearity (VIF > 35).

Heteroscedasticity & Autocorrelation Checks

- **Breusch-Pagan & White Tests:** Confirmed heteroscedasticity issues.
- **Durbin-Watson Statistic:** Determined whether autocorrelation was present.
- **Log Transformation & Box-Cox:** Attempted to stabilize variance in % price change, though heteroscedasticity remained.

When I realized that I am having trouble with solving heteroscedasticity, I figured I would get more data from the year 2019 to see whether the problem goes away.

Model	2019
Multiple Linear Regression (MLR)	1.000
Ridge Regression	0.9999
Lasso Regression	0.9999
Elastic Net Regression	0.9999

Support Vector Regression (SVR)	0.917
Random Forest Regression	0.9984
Decision Tree Regression	0.9995
K-Nearest Neighbors (KNN)	0.885
Weighted Least Squares (WLS)	1
Logistic Regression (Logit)	Singular Matrix

I added 1,000 more data points for 2019. After combining variables, multicollinearity and heteroscedasticity persisted.

VIF Factor	Features
3.432	const
NaN	Ticker
∞	beta
∞	EPS (2019)
1.245	Gross Margin (2019)
1.284	+/- price change
1.021	Returns (2019)
1.033	Liquidity (2019)
1.046	Efficiency (2019)

- Breusch-Pagan test p-value: 1.54e-101 → Strong evidence of heteroscedasticity.

- Durbin-Watson statistic: 2.0002 → No evidence of autocorrelation.

This result was actually worse than the previous dataset with only 500 observations. To address multicollinearity and heteroscedasticity, I tried removing beta and EPS (2019). However, I found that the model becomes extremely weak when either is removed in isolation. I applied Weighted Least Squares (WLS) estimation to address heteroscedasticity, which improved model stability, but both multicollinearity and heteroscedasticity still persisted. Interestingly, EPS (2019) emerged as the most important predictor after this adjustment.

To quantify the impact of removing each variable, I ran Ridge Regression and tracked the adjusted R-squared:

Variable Removed	Adjusted R-squared	Effect
Beta	0.9999999999991456	Fixes multicollinearity; heteroscedasticity remains
EPS (2019)	0.015872706571313788	Model becomes extremely weak
Gross Margin (2019)	0.9999999999991485	Model remains strong
Returns (2019)	0.9999999999991488	Model remains strong
Liquidity (2019)	0.9999999999991287	Model remains strong
Efficiency (2019)	0.9999999999991454	Model remains strong

After removing only beta, multicollinearity was resolved, but heteroscedasticity remained. When I also removed EPS (2019), both multicollinearity and heteroscedasticity were resolved, and the model still maintained a very high adjusted R-squared after running Ridge regression on it.

Variable	VIF
Gross Margin (2019)	1.136
Returns (2019)	1.018
Liquidity (2019)	1.133

Efficiency (2019)	1.042

Test	Result
Breusch-Pagan Test	p-value = **0.9467** → No heteroscedasticity
Durbin-Watson Statistic	**2.0000** → No autocorrelation

With both beta and EPS (2019) removed, the final model shows no signs of multicollinearity, no heteroscedasticity, no autocorrelation, and still boasts an extremely strong adjusted R-squared of 0.9999999999991485.

Reflection

This project was a valuable and challenging experience that deepened my understanding of both financial data and predictive modeling techniques. One of the most successful aspects was the thorough preprocessing and diagnostic testing of the data. By identifying and addressing multicollinearity, heteroscedasticity, and autocorrelation, I was able to build models that were not only statistically sound but also interpretable and realistic. Techniques like variable transformation and the application of regularization models (e.g., Ridge, Lasso, and Elastic Net) allowed for improved model stability and reduced overfitting, especially in the presence of highly skewed predictors. However, not everything went smoothly. One major difficulty was handling extreme skewness and outliers in key financial ratios such as EPS and ROE. While removing these outliers could have improved model performance, doing so would have meant excluding some of the most informative and economically significant data points. Striking the right balance between robustness and representativeness was a recurring challenge. Additionally, integrating and cleaning financial data from multiple sources required far more time than expected, often delaying the modeling phase.

In hindsight, I would allocate more time to data exploration and feature engineering, as well as consider dimensionality reduction techniques to manage high collinearity among financial variables. I also would have experimented with more ensemble models earlier in the process, as methods like Random Forest and Gradient Boosting offered strong performance with relatively little tuning compared to linear models. To others undertaking similar work, I would recommend investing substantial effort upfront in understanding the data—particularly how financial ratios behave across different industries and timeframes. Also, don’t rely solely on traditional metrics like

R^2 or MSE; interpretability and economic reasoning should guide model evaluation as much as raw accuracy. Lastly, document everything. With so many iterations and diagnostics, clear notes and organized code are essential to stay on track. Most importantly, I learned that financial data science is as much about judgment and trade-offs as it is about technical skill. Real-world data is rarely clean or well-behaved, and successful modeling often means embracing imperfection while making informed, transparent decisions. This project gave me practical exposure to navigating that complexity and highlighted the importance of combining domain knowledge with statistical rigor.

Conclusion & Future Work

This project explored the use of financial ratios to predict year-ahead stock price changes across a diverse set of companies, leveraging a range of linear, regularized, and ensemble regression models. The analysis revealed that while financial datasets are often plagued by skewness, outliers, and multicollinearity, these issues can be addressed with rigorous diagnostic testing and thoughtful model selection. Regularized models like Ridge, Lasso, and Elastic Net effectively handled overfitting and high dimensionality, while ensemble methods such as Random Forests and Support Vector Regression demonstrated strong predictive performance with limited tuning. Ultimately, the project succeeded in balancing predictive accuracy with model interpretability—an essential trade-off in financial analytics.

Looking forward, several promising paths could be pursued to extend this work. One clear direction is the integration of macroeconomic indicators—such as interest rates, inflation expectations, or GDP growth—which often influence market sentiment and investment flows. Including such variables could provide a broader contextual backdrop for firm-level financials, enhancing predictive power. Another avenue is the incorporation of forward-looking and alternative data sources, including analyst earnings forecasts, credit ratings, ESG scores, or even textual sentiment extracted from financial news and earnings call transcripts using natural language processing (NLP). These variables may capture investor expectations and qualitative factors not reflected in historical ratios. Expanding the scope of the dataset both temporally and geographically could also yield valuable insights. Including data from post-2023 periods or from international markets would allow for robustness checks across different economic regimes and regulatory environments, possibly revealing model limitations or opportunities for generalization.

From a methodological standpoint, exploring deep learning approaches (e.g., recurrent neural networks for time series data or transformer-based architectures for unstructured inputs) could open new possibilities, especially

if the dataset is enriched with sequential or textual features. Similarly, experimenting with hybrid models that blend machine learning with traditional econometric structures might combine the strengths of both paradigms—maintaining interpretability while capturing nonlinearity and interactions. Finally, building a real-time or dynamic modeling system that updates predictions as new data arrives could make the framework more applicable in practice. Such a system might use rolling windows, online learning algorithms, or Bayesian updating to remain adaptive in fast-changing markets. In conclusion, this project lays a strong foundation for stock return prediction using financial fundamentals, but it also highlights the importance of continuously evolving the methodology to keep pace with financial markets' complexity and data richness.

Works Cited

Bloomberg. (n.d.). *Company profiles*. Bloomberg. Retrieved from https://www.bloomberg.com

Finsheet. (n.d.). *Financial data analysis tools*. Finsheet. Retrieved from https://www.finsheet.io