Report on Capstone 3: Stock Portfolio Optimization

Springboard Data Science Bootcamp Cohort May 2023
Steve Walters
Prepared April 25, 2024

Introduction

Problem Statement:

This analysis aims to identify and optimize a portfolio of stocks from the S&P 500 using Efficient Frontier, Black-Littman and Hierarchical Risk Parity (HRP). One of the three models will be chosen as superior in order to perform additional regression analysis to determine the predictive capabilities of the optimization. We will further evaluate its forward-looking performance against market benchmarks.

Significance:

Understanding the efficiency and robustness of the chosen portfolio optimization method could significantly benefit portfolio management by maximizing returns relative to risk.

Data Description and Preprocessing

Data Sources:

Stock price data was sourced from Yahoo Finance for all S&P 500 constituents listed in a CSV file.
Stock symbol data was sourced from Github as constituents.csv

Data Cleaning:

Data for each ticker was downloaded and checked for completeness. Tickers with missing data were identified and excluded.

Feature Engineering:

Features such as daily returns and volatility were computed. Dummy variables were not explicitly mentioned but are implied in categorical analyses like GICS sectors.

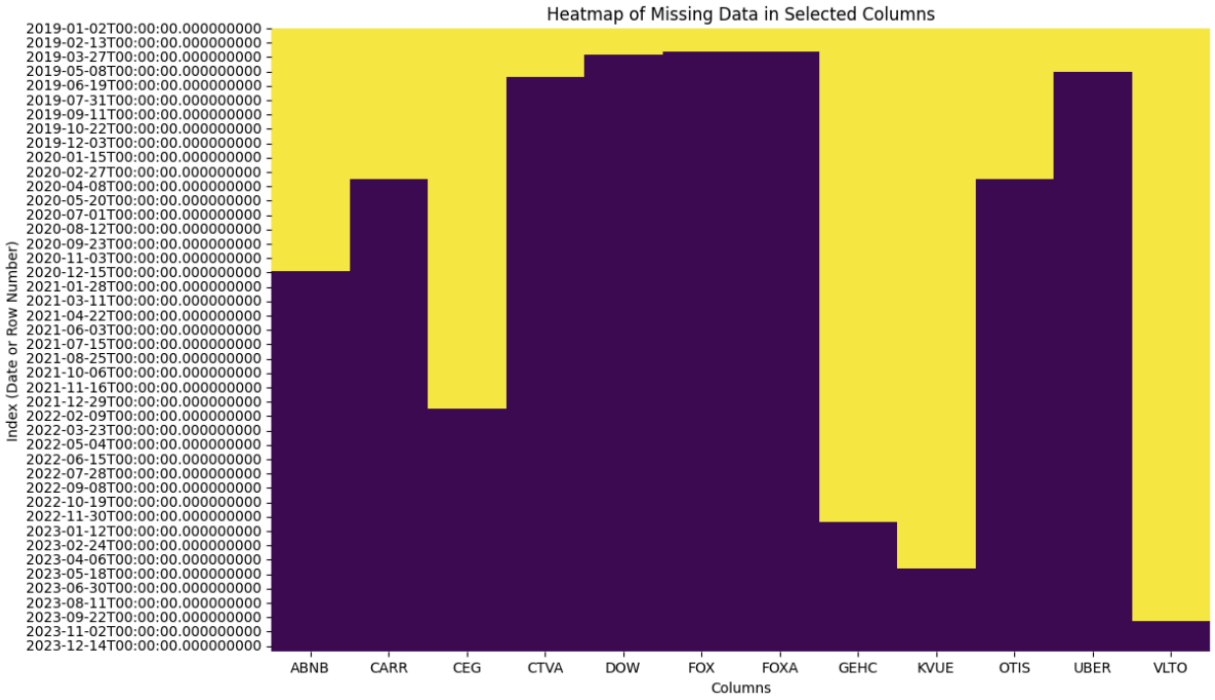
Standardization:

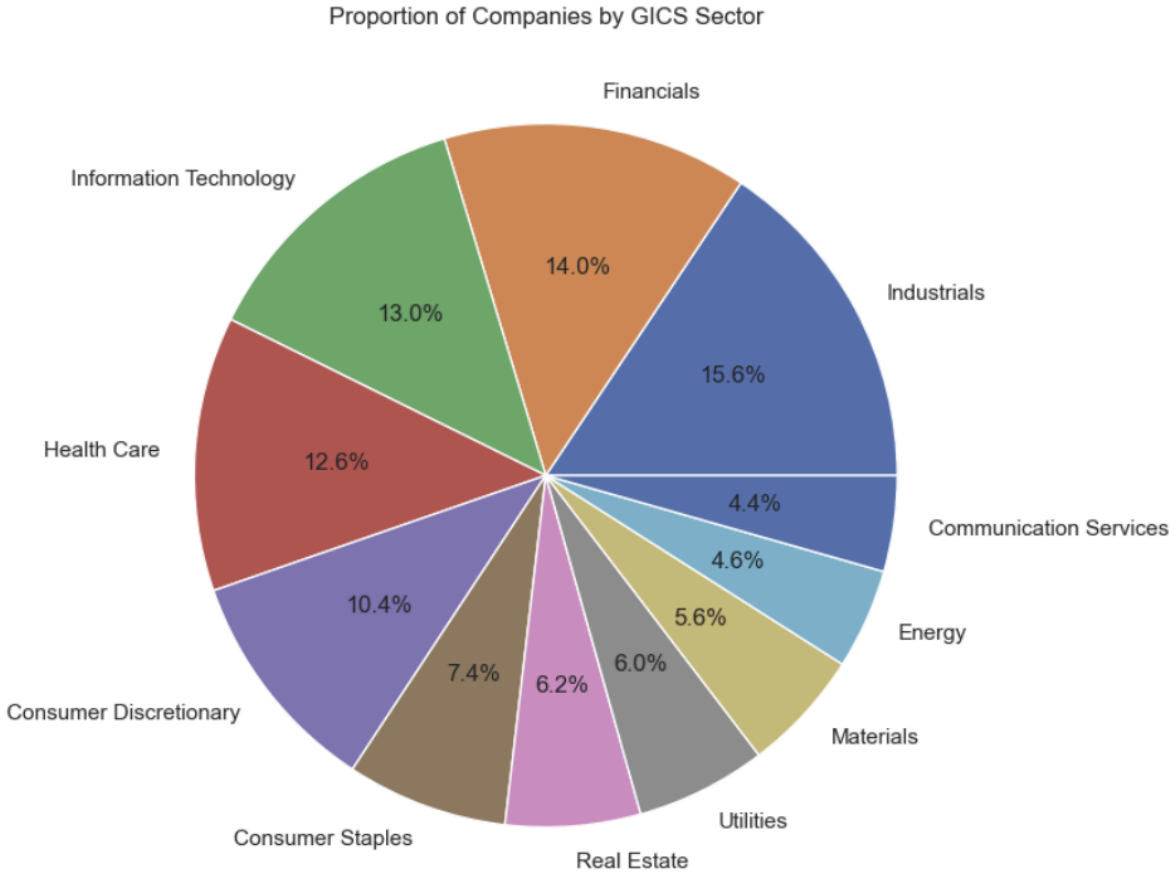
Data features were magnitude standardized for modeling purposes, ensuring that scale discrepancies do not skew results.

Exploratory Data Analysis (EDA)

Visualizations:

Various plots including heatmaps for missing data, Z-score distributions to identify outliers, and sectoral distributions were utilized to understand data characteristics and relationships.


Heatmap of Missing Data in Selected Columns
Index (Date or Row Number)
2019-01-02T00:00:00.000000000
2019-02-13T00:00:00.000000000
2019-03-27T00:00:00.000000000
2019-05-08T00:00:00.000000000
2019-06-19T00:00:00.000000000
2019-07-31T00:00:00.000000000
2019-09-11T00:00:00.000000000
2019-10-22T00:00:00.000000000
2019-12-03T00:00:00.000000000
2020-01-15T00:00:00.000000000
2020-02-27T00:00:00.000000000
2020-04-08T00:00:00.000000000
2020-05-20T00:00:00.000000000
2020-07-01T00:00:00.000000000
2020-08-12T00:00:00.000000000
2020-09-23T00:00:00.000000000
2020-11-03T00:00:00.000000000
2020-12-15T00:00:00.000000000
2021-01-28T00:00:00.000000000
2021-03-11T00:00:00.000000000
2021-04-22T00:00:00.000000000
2021-06-03T00:00:00.000000000
2021-07-15T00:00:00.000000000
2021-08-25T00:00:00.000000000
2021-10-06T00:00:00.000000000
2021-11-16T00:00:00.000000000
2021-12-29T00:00:00.000000000
2022-02-09T00:00:00.000000000
2022-03-23T00:00:00.000000000
2022-05-04T00:00:00.000000000
2022-06-15T00:00:00.000000000
2022-07-28T00:00:00.000000000
2022-09-08T00:00:00.000000000
2022-10-19T00:00:00.000000000
2022-11-30T00:00:00.000000000
2023-01-12T00:00:00.000000000
2023-02-24T00:00:00.000000000
2023-04-06T00:00:00.000000000
2023-05-18T00:00:00.000000000
2023-06-30T00:00:00.000000000
2023-08-11T00:00:00.000000000
2023-09-22T00:00:00.000000000
2023-11-02T00:00:00.000000000
2023-12-14T00:00:00.000000000
ABNB
CARR
CEG
CTVA
DOW
FOX
FOXA
GEHC
KVUE
OTIS
UBER
VLTO
Columns


Proportion of Companies by GICS Sector
Financials
14.0%
Industrials
15.6%
Communication Services
4.4%
Energy
4.6%
Materials
5.6%
Utilities
6.0%
Real Estate
6.2%
Consumer Staples
7.4%
Consumer Discretionary
10.4%
Health Care
12.6%
Information Technology
13.0%

Model Development and Evaluation

Optimal Portfolio:

Three methods were examined to determine the optimal portfolio of ten stocks from the S&P 500. These stocks were chosen based on the top performing stocks over the time period used (1/1/2019 through 1/1/2024) and were limited to no more than two stocks from any single GICS sector in order to maintain diversity in the portfolio. The ticker symbols and GICS sectors used in the optimization were: ENPH (Information Technology), SMCI (Information Technology), TSLA (Consumer Discretionary), DECK (Consumer Discretionary), BLDR (Industrials), CARR (Industrials), MRNA (Health Care), DXCM (Health Care), CEG (Utilities), PCG (Utilities).

After performing portfolio optimization using the three models – Efficient Frontier, Black-Litterman, and Hierarchical Risk Parity (HRP) – it was determined that the HRP optimization gave the best risk/reward and Sharpe Ratio. Regression analysis that follows was focused on using the optimized portfolio weights from this HRP model.

Train-Test Split:

The dataset was split into training and testing subsets to evaluate the models' performance objectively.

Model Building:

Three models were built: Linear Regression, Random Forest, and XGBoost, chosen for their diverse assumptions and approaches to regression. Lagging was used in the creation of the models to create autoregressive models, which use past values of the target variable to predict future values.

Performance Comparison:

Comparison tables were created to evaluate MSE and R^2 scores across models, highlighting their predictive accuracies.

5 Lag Results

	MSE	R-Squared
Linear Regression	0.00046391218020657885	-0.006713751228128784
Random Forest	0.0005023298867902183	-0.09008218852842798
XGBoost	0.000505188501737775	-0.0962855328249499

Optimized 14 Lag Results

	MSE	R-Squared
Linear Regression	0.0004498718816739824	0.01782609351580966
Random Forest	0.0004988837400407312	-0.08917807890989593
XGBoost	0.0004710748447453201	-0.02846485712401603

Final Model Selection and Application

Selection:

The XGBoost model was initially identified as the final model due to its balance between complexity and performance. However, further optimization of the lag period identified the linear regression model as the ideal final model as can be seen in the table above.

Application and Results Review:

Further enhancements were made to the modeling approach by adjusting the number of lagged features used in the analysis. Initially, 5 lags were utilized based on standard practice; however, further exploration using ACF and PACF identified that extending the lag to 14 provided a more accurate reflection of temporal dependencies in the data. This adjustment led to a reevaluation of the models' performances. The Linear Regression model, when reapplied with the optimal 14 lags, showed an unexpected improvement in performance, surpassing the XGBoost model that was previously favored. This change was quantified by a lower MSE and a positive R^2 score, indicating both high accuracy and a reasonable explanation of variance through the model.

Given the new findings, the Linear Regression model with 14 lags was selected as the final model. This decision was based on the superior MSE of 0.00044987188167398243 and an R^2 of 0.01782609351580966, demonstrating not only improved prediction accuracy but also an ability to explain more variability in the data compared to the baseline model. This outcome highlights the importance of model reevaluation and feature adjustment in data science workflows, particularly in complex domains such as financial markets where model performance can significantly impact decision-making.

Results and Discussion

Findings:

The HRP-based portfolio demonstrated superior risk-adjusted returns compared to traditional methods during the historical period. Forward-looking analysis against the S&P 500 using data from 2024 was also performed and showed promising results.

Comparison of Forward-Looking Results

	Optimized HRP Portfolio	S&P 500 Benchmark
Expected Annual Return	141.93%	22.55%
Annual Volatility	2.52%	11.66%
Sharpe Ratio	56.3763	1.4862

Conclusion

Summary:

The analysis confirmed the effectiveness of the HRP method in constructing a well-diversified portfolio that outperforms traditional models in terms of Sharpe ratio and overall risk management. The combination of the results from a 14 period lagged linear regression model and the actual comparison of forward looking results using the HRP optimized portfolio indicates some promise in the predictive power of the model. In the forward-looking analysis the HRP optimization outperformed the S&P 500 by more than 6x, while also reducing volatility and greatly increasing the Sharpe Ratio.

Future Work:

Further research could explore deeper integration with machine learning techniques to refine asset selection and weight allocation. The model could also be modified to allow for user input of custom portfolios and regular updates to portfolio weights.