Equity Investment Using Python: From Data Mining to Portfolio Strategy Formulation

Prepared by : Chulwon Chae

Date : April 13, 2024

Executive Summary: This personal project aims to design and manage a diversified stock portfolio using Python, focusing on extracting and analyzing financial data from various online sources. The core objective is to employ key financial metrics such as value, momentum, and quality, integrated into a sophisticated multi-factor portfolio strategy. Data for the project is gathered via systematic web crawling from sources like Naver Finance, the Korean Exchange, and Yahoo Finance, and consolidated in a MySQL database. The methodology combines qualitative and quantitative measures, including financial statement analysis and technical analysis via TA-Lib, to select and optimize the portfolio. Notably, the investment strategy employs a rigorous z-score computation for balanced sector exposure, supplemented by backtesting through the Backtrader package.

Introduction

Objective : The objective of this project is to develop a stock portfolio using Python. This initiative is a personal project and is not intended to serve as investment advice or any academic purpose. All investment decisions made based on this project are solely the responsibility of the individual.

In this project, I will leverage Python to extract data on stock ticker, prices, and financial statements from diverse online sources, consolidating this information into an MySQL database. My aim is to utilize key financial metrics such as value, momentum, and quality to strategically select stocks. I will then employ sophisticated portfolio strategies to design and manage diversified investment portfolios.

Market Analysis :

It's critical to analyze the current market conditions, characterized by an inverted yield curve persisting for approximately two years. Historically, such inversions often precede economic recessions and warrant a cautious approach to investment. Following COVID-19, the Federal Reserve's quantitative easing measures have significantly increased liquidity in the economy. Despite this influx, the stock market has not experienced a dramatic downturn, suggesting that the effects of these policies are still buffering the market. In times like these, relying on comprehensive market analysis and data-driven strategies is essential to navigate potential volatility and capitalize on opportunities in a prudent manner.

Data and Methodology

Data Source :

The data was meticulously collated through the systematic crawling of publicly available online resources, culminating in the creation of a database within MySQL Workbench. The crawling process necessitated the use of both GET and POST requests, for which the BeautifulSoup and Selenium WebDriver libraries were employed. The coding phase of the crawling operation was particularly challenging and demanded a considerable degree of trial and error to perfect.

Data	Source	Crawling Method
Updated Date, Korean Stock Price	Naver Finance (ex.Yahoo Fin)	GET
Korean Stock List, PER/PBR	Korean Exchange (KRX)	POST
Korean Stock Sector Info, Financial Statement	FN Guide	GET
Global Stock List	Investing.com	GET
Global Stock Price	yfinance	Python Library
Global Stock Financial Statement	yahooquery	Python Library

Fig 1.Data and Source

Methodology :

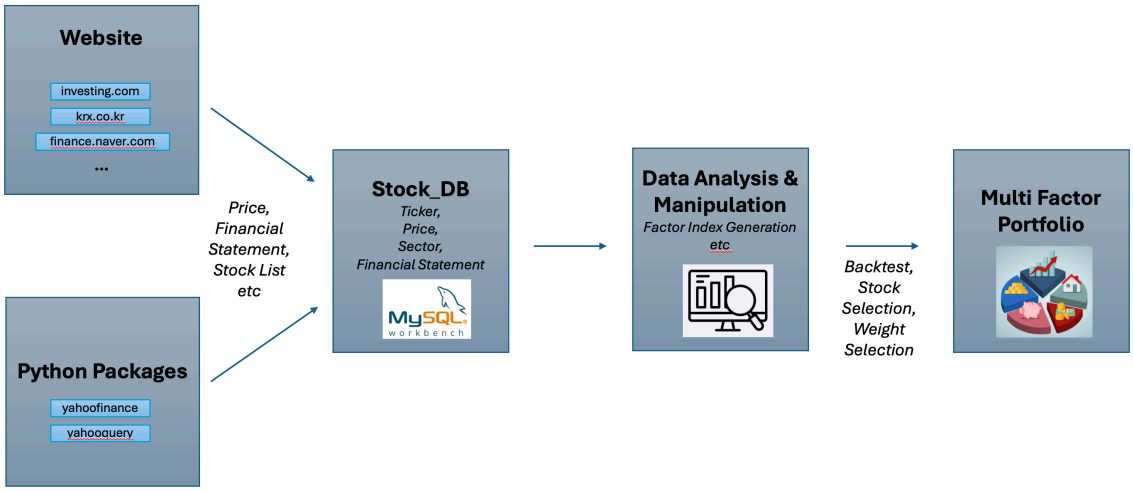

Fig 2. Data Flow

Investment Strategy

Multi-Factor Portfolio

This strategy harnesses the synergies of quality, value, and momentum metrics to meticulously select the top 20 performing stocks. In crafting the quality component, critical financial dimensions such as net income, gross profit, cash flows from operating activities, equity, and

total assets were extracted from comprehensive financial statement databases. The valuation aspect of the strategy filters out securities with negative metrics to ensure a focus on fundamentally sound investments. The momentum feature is engineered using the 12-month logarithmic returns to capture the persistence of performance, further refined by employing the K-Ratio as a measure of return consistency for ranking purposes. To consolidate these facets into a cohesive strategy, I computed the z-scores for each factor, assigning equal weighting to achieve a balanced exposure across all sectors. Outliers were excluded from the analysis by trimming the dataset at the 1st and 99th percentiles during the z-score computation process. This rigorous analytical process culminates in the final selection of the prime 20 stocks poised for inclusion in the investment portfolio.

Factors	Features (Index)
Quality	Return on Equity (ROE), Gross Profitability (GPA), Cash Flow from Operations (CFO)
Value	Price-to-Earnings Ratio (PER), Price-to-Book Ratio (PBR), Price-to-Sales Ratio (PSR), Price-to-Cash Flow Ratio (PCR), Dividend Yield (DY)
Momentum	12-month Returns, K-Ratio

Fig 3. Factor investing strategies and index used

	Ticker	Stock_Name	Sector	ROE	GPA	CFO	DY	PBR	PCR	PER	PSR	12M	K_ratio	z_quality	z_value	z_momentum	qvm	invest
15	000270	기아	경기관련소비재	0.0548	0.2885	0.1502	0.0363	0.9237	3.3112	4.2110	0.3925	0.6420	35.2640	-3.9913	-3.0605	-2.9579	-1.1761	Y
17	000320	노루홀딩스	산업재	0.0204	0.2364	0.0733	0.0421	0.2354	1.8066	2.8866	0.1313	0.1072	19.2504	-2.5229	-6.8638	-1.9310	-1.1670	Y
96	002030	아세아	소재	0.0234	0.1456	0.0815	0.0220	0.1918	1.3941	2.0462	0.1716	0.4330	57.0209	-2.3529	-6.2673	-3.0708	-1.2713	Y
141	003030	세아제강지주	소재	0.0490	0.2360	0.1282	0.0097	0.4357	1.9129	2.2228	0.2409	0.4953	21.9700	-4.6214	-3.8994	-2.6019	-1.2751	Y
209	004710	한솔테크닉스	IT	0.0292	0.1873	0.1015	0.0143	0.4858	2.0879	4.1642	0.1337	0.3094	24.9418	-2.6290	-6.5708	-1.8181	-1.1366	Y
373	009970	영원무역홀딩스	경기관련소비재	0.0478	0.3178	0.1071	0.0382	0.2551	1.7095	1.3333	0.2394	0.2540	55.8490	-3.5463	-6.5758	-2.7594	-1.4017	Y
415	011560	세보엠이씨	산업재	0.0354	0.1479	0.1877	0.0547	0.5048	1.2614	3.5646	0.1025	0.1220	24.0508	-3.3962	-6.7155	-2.1736	-1.3030	Y
515	017670	SK텔레콤	커뮤니케이션서비스	0.0240	0.5728	0.1673	0.0664	0.8889	2.1441	9.2554	0.6261	0.0394	4.2185	-3.7899	-5.8630	-1.8202	-1.2215	Y
700	033270	유나이티드제약	건강관리	0.0289	0.3539	0.1363	0.0159	1.1553	6.7442	9.9846	1.4789	0.1638	12.1150	-4.1753	-4.3212	-1.7550	-1.1228	Y
835	041830	인바디	건강관리	0.0385	0.5225	0.1450	0.0115	1.5612	9.5566	10.1268	2.0773	0.2675	18.7377	-4.8072	-2.5136	-2.2826	-1.1250	Y
988	054040	한국컴퓨터	IT	0.0350	0.2001	0.2315	0.0450	0.6880	2.4282	4.9097	0.2216	0.2138	15.9336	-3.6375	-7.3962	-1.1261	-1.2297	Y
1000	054670	대한뉴팜	건강관리	0.0451	0.5291	0.1293	0.0111	1.1734	5.6790	6.5066	0.6356	0.0272	5.6258	-4.8190	-4.9463	-1.0897	-1.1558	Y
1228	079940	가비아	IT	0.0305	0.3195	0.1822	0.0043	0.8494	3.1900	6.9582	0.8725	0.6483	47.2577	-4.1913	-2.8798	-2.7957	-1.1599	Y
1230	079960	동양이엔피	IT	0.0306	0.1570	0.1510	0.0150	0.5321	2.5835	4.3534	0.2948	0.3867	44.8006	-2.8200	-6.2787	-2.4475	-1.2338	Y
1266	084010	대한제강	소재	0.0302	0.1873	0.1151	0.0588	0.3867	2.3414	3.2065	0.1939	0.1383	26.3457	-3.7784	-6.4987	-2.2842	-1.3496	Y
1380	094970	제이엠티	IT	0.0676	0.1660	0.2511	0.0368	0.6234	1.9956	2.3057	0.8661	0.2846	21.1500	-3.6879	-6.6704	-1.6293	-1.2505	Y
1549	123040	엠에스오토텍	경기관련소비재	0.0606	0.2092	0.1762	0.0093	0.4602	0.8499	1.8980	0.1076	0.2014	22.8595	-3.7625	-4.6554	-2.0745	-1.1503	Y
1551	123410	코리아에프티	경기관련소비재	0.0346	0.2210	0.1534	0.0227	0.6013	1.5501	4.3486	0.1551	0.4626	31.7150	-3.2381	-4.5939	-2.7792	-1.1889	Y
1869	224110	에이텍모빌리티	IT	0.0293	0.3024	0.2092	0.0066	0.8806	2.5020	7.5059	0.4427	0.4909	42.7580	-4.1579	-4.0148	-2.5958	-1.2255	Y
1931	243070	휴온스	건강관리	0.0366	0.5939	0.0644	0.0158	1.5004	14.7801	10.2529	0.8491	0.2821	29.4911	-4.1753	-3.4116	-2.6268	-1.1788	Y

Fig 4. Finalized top 20 selected stock list

Technical Analysis strategy

The investment approach using technical analysis can be easily implemented using the TA-Lib package. TA-Lib, or Technical Analysis Library, is a widely-used open-source software library that provides tools for financial market analysis and algorithmic trading. It is particularly well-known for its comprehensive collection of indicators, which are fundamental to conducting technical analysis in trading.

Portfolio Optimization

The selection of individual stock weights for portfolio optimization can be efficiently addressed using the Riskfolio-Lib library. This library facilitates the construction of traditional portfolio types such as Minimum Risk, Maximum Return, Utility, or Sharpe Ratio portfolios. Additionally, it allows for the imposition of constraints on investment weights for different asset classes or individual securities.

Back testing

The backtesting process can be easily run using the Backtrader package. Backtrader is a popular Python library designed for backtesting and trading algorithmic strategies. It provides a robust framework that allows users to test their trading ideas using historical data.

Future Work

To enhance the portfolio's performance and adaptability, below are the future initiatives for this project, on which I will continue to work;

1. US Market Expansion: Explore stock selection in the U.S. market to diversify and potentially enhance the portfolio's return profile. This expansion requires adapting the data crawling techniques to sources relevant to U.S. financial markets.
2. Risk Control Strategies: Implement advanced risk control strategies such as stop loss and trailing stop mechanisms. These strategies will help in mitigating risks during significant market downturns and protect the portfolio against large, unexpected losses.
3. Further Analytical Enhancements: Integrate additional analytical tools and strategies, such as machine learning algorithms for predictive analytics and sentiment analysis of market trends from news sources and social media. These tools can provide deeper insights into market dynamics and help refine stock selection and timing decisions.

Conclusion

The project successfully demonstrates the feasibility of using Python to develop a robust investment strategy grounded in multi-factor analysis. The use of sophisticated data crawling techniques, combined with advanced financial libraries, facilitated a data-driven approach to stock selection. The portfolio, optimized for a balance of quality, value, and momentum factors, showed potential resilience and performance in the volatile market conditions post-COVID-19. The technical infrastructure, including MySQL and various Python libraries, proved effective in handling complex data and analytical tasks, thus reinforcing the viability of technology-driven personal investment strategies.

If you have any questions regarding the research and code, please do not hesitate to contact me via email at chulwonchae@gmail.com.