

**Predicting Institutional Investment Trends
with Machine Learning:
Emulating Berkshire Hathaway's
Long-Term Portfolio Strategy**

Vivian Cheng

Content

Introduction.....	3
Non-machine learning approaches.....	5
Machine learning.....	7
Current gaps in studies.....	9
Proposal.....	9
Methodology.....	11
Results.....	13
Discussion.....	13
Conclusion.....	17
Reference.....	18

Introduction

Institutional trading, such as hedge funds, mutual funds, pension funds, and holding companies, plays a vital role in the global financial market. With large amounts of capital and sophisticated investment strategies, the impact on the market is often more significant than retail investors, by influencing stock prices and market trends through their buying and selling decisions. Among these institutional investors, hedge funds focus on short-term and aggressive strategies to generate high returns, using complex techniques like leverage, derivatives, and short-selling, which are highly dynamic but risky, hence difficult to replicate and follow. Whilst mutual funds adopt a more conservative approach, favouring stability over high returns with long-term, diversified portfolios to reduce risk. Unlike both of them, holding companies like Berkshire Hathaway stand out amongst institutional investors for their value investment approach, emphasising a long-term, concentrated and relatively stable portfolio, which is able to ensure high return and stability through institutional trading.

Driven by Warren Buffett's value-investing philosophy, Berkshire Hathaway has maintained its consistently superior performance through a carefully curated, long-term portfolio. Buffett's strategy focuses on large-cap growth stocks (Martin & Puthenpurackal, 2005), which are large companies with significant growth potential, rather than the traditional value stocks, which are typically undervalued stocks with growth potential. Berkshire minimizes downside risks by carefully selecting companies based on key factors such as long-term performance, measured by return on equity (ROE) over the past 5 to 10 years, and financial stability, indicated by a favourable debt-to-equity (D/E) ratio that prioritizes equity over debt (Team 2024). Additionally, the company favours businesses with steadily increasing profit margins, limited reliance on commodities, and a trustworthy reputation built through long-term public presence. Buffett's approach also emphasizes investing in companies whose intrinsic value is below their current market worth.

Berkshire further enhances its value by reinvesting dividends and utilizing the insurance float (McFarlane 2024), the temporary holding of policyholder premiums before payouts, to fund investments, reducing reliance on debt. The company's

low-leverage strategy reduces risk while avoiding stock splits to maintain high share prices, which attracts long-term, committed investors. Additionally, Berkshire's portfolio is highly concentrated, with the top five holdings making up an average of 73% of its total value, making it ideal for tracking and replication purposes. With this carefully constructed portfolio, Berkshire outperformed the S&P 500 Index by an average of 11.14% annually, exceeded a value-weighted index of all stocks by 10.92% per year, and outpaced a Fama and French characteristic-based portfolio by 8.56% per year, delivering consistent success in 27 out of 31 years. Berkshire's ability to generate high returns with relatively low risk makes it a vital example for studying institutional trading strategies and developing predictive machine learning models on top of them.

Non-machine learning approaches

Historically, traditional stock market prediction is based on two main approaches - fundamental and technical analysis. Fundamental analysis, a relatively long-term approach famously valued by Buffett, focuses on a top-down approach and utilises publicly available information like financial statements. It is also seen by many fund managers to be more reasonable as it takes quantitative references from the market capitalization-to-GDP ratio (*Stock market prediction 2024*), which indicates the relative intrinsic value, and compares it with the current market value. Hence, the approach targets underestimated stocks with long-term potential by analysing their true value with existing information like financial statements and current market conditions. As mentioned previously, holding companies like Berkshire under the influence of Buffett, rely on this approach with value-investing techniques, selecting underestimated companies with long-term financial potential based on ROE, D/E ratio and other indicators (Team, Warren Buffett's investment strategy).

On the other hand, technical analysis is often seen to be more short-term which is often favoured by hedge funds and short-term investors, opting for the utilisation of time series analysis, which forecasts price direction based on historical data, primarily price and volume (*Stock market prediction 2024*). In specific, instead of focusing on a company's fundamentals, technical analysts believe that stock prices reflect all the information they need and that history tends to repeat based on market psychology. They thus focus on working with charts with patterns like head-and-shoulder, and the famous candle stick pattern (*Stock market prediction 2024*), with techniques like exponential moving average (EMA), oscillators, support and resistance levels or momentum and volume indicators (*Stock market prediction 2024*), to provide them with adequate evidence of future price movement that supports their trading decisions.

Whilst these two approaches are regarded as pillars of stock predictions, we should not disregard their limitations. Fundamental analysis, despite its evaluation based on publicly available and quantitative data like financial statements, is largely driven by subjective interpretation and biased judgment to determine a company's true value and long-term potential. To achieve a more accurate evaluation, in-depth data

analysis is time-consuming and may not apply in short-term forecasting. Data reliability and accuracy are often challenged in this fast-paced society where market conditions change rapidly, and not all necessary information may be available. Whereas technical analysis is often criticised for its simplified assumptions that past data can solely determine future price movement. The efficient-market hypothesis (EMH) (*Stock market prediction 2024*) challenges this idea by claiming that market price is also dependent on new, unpredictable information, which cannot be anticipated solely by studying the historical market trend. On the whole, both fundamental and technical analysis approaches are inefficient in processing large sets of information and acknowledging subtle relationships between data, hence a more advanced strategy is highly sought after.

Machine learning

With traditional approaches, like fundamental and technical analysis, struggling to keep pace in the fast-evolving financial environment, while handling massive datasets and failing to identify less prominent aspects of information, the development of machine learning (ML) has quickly risen in the relevant field to address these challenges and improve prediction accuracy and efficiency. ML algorithms are said to outperform traditional methods in financial market forecasting (Ryll & Seidens, 2019), utilising initial techniques including artificial neural networks (ANN) and random forests (RF). These models have shown their capability in handling datasets with volume and complexities and identifying subtle patterns for better prediction performance.

In particular, ANN offers both feedforward with backpropagation of error approach and time recurrent neural network approach (RNN). Whilst the feedforward approach focuses on static information, RNN generally has a better performance in terms of financial forecasting as it better captures sequential dependencies in data (Ryll & Seidens, 2019), where past values influence future predictions, revealing “exploitable temporal patterns” (Ryll & Seidens, 2019). Specifically, RNN processes data across multiple time stamps and utilises a joint approach for forecasting different time horizons. Interconnected data from different time horizons can provide a more holistic and comprehensive forecast as it takes reference from different time predictions.

However, the increase in parameters also increases model complexity and imposes risks of overfitting (*Stock market prediction* 2024), with a larger impact of error due to the interdependence of time horizons that may propagate and affect each other, which potentially deteriorates the applicability and accuracy of the ML model. To counter this, there has been an increase in advocacy recently for an ensemble of independent ANNs methods instead, allowing multiple independent ANNs to specialise for a target function like future lows and highs (*Stock market prediction* 2024), hence reducing complexity, allowing errors isolation and improving accuracy.

Overall, ML is applicable to improve both technical and fundamental analysis in areas of price prediction and value estimation. In technical and statistical arbitrage, ML helps identify and exploit mispricing with efficient detection of relationships between variables in datasets. On the other hand, ML also improves fundamental investing by identifying long-term potential and values through sentimental analysis including public and media response to markets. Whilst past price alone only contributes to limited predictability (Jiao, Jakubowicz 2017), incorporating external information and events helps improve performance in event-driven trading. Moreover, alternate data trends including search volumes in Google trends, views of relevant wiki pages or even news feeding with headlines have become more popular recently, with the incorporation of text mining and ML models, to show how external response is relevant to internal price movements.

Multiple machine learning techniques, namely supervised learning, unsupervised learning and reinforcement learning, contribute to the success of ML stock prediction. Supervised learning involves training with labelled data with regression and classification techniques, to predict continuous values like stock prices or discrete labels like whether a stock will go up or down (Snow 2019), which helps use historical data to predict future movement. Unsupervised learning aims to identify hidden patterns in vast and unlabelled datasets. It utilises methods like K-means clustering and Principal Component Analysis (PCA) for feature grouping and reducing dimensions of the dataset to reduce complexity and potential overfitting (Snow 2019), allowing the discovery of subtle relationships amongst features that are less prominent in traditional methods. Reinforcement learning focuses on learning through feedback, and is effective in model optimisation, in relevance to event-driven trading and later stages of model development. All of these techniques contribute to a more holistic and comprehensive development of ML algorithms for better prediction experience.

Current gaps in studies

Current ML models mostly focus on the technical and short-term aspects of stock price predictions that typically utilise techniques like an individual investor. While there is a difference in investing behaviours between the individual and institutional investors, the former tend to sell more while the latter tend to buy more when the stock market is better performing (Griffin, Harris, Topaloglu 2003), it is less common with a machine learning model to replicate institutional trading, which incorporates long-term value investing, and creates a gap in replicating institutional trading behaviours. There is also insufficient focus on using fundamental metrics in ML models as most focus on high-frequency price prediction derived from historical price data, without comprehensive contribution of external factors. Specifically, public filings like SEC Form 13F, in which institutional investment manager disclose their investment portfolio, are not utilised adequately in current ML models. It may be argued that there is more flexibility in evaluating each company independently for more customisable portfolio purposes, however it is refuted by the incomparable ability institutional trading firms have in terms of the scope of information accessible for evaluation through their connections, compared to individual researchers like us whose best source of evaluation may be based on ML and AI. Thus, utilising these strategic data that provide valuable insights that should be equally valued as historical prices, we can bridge the gap between short-term stock price prediction and long-term institutional value investing replication.

Proposal

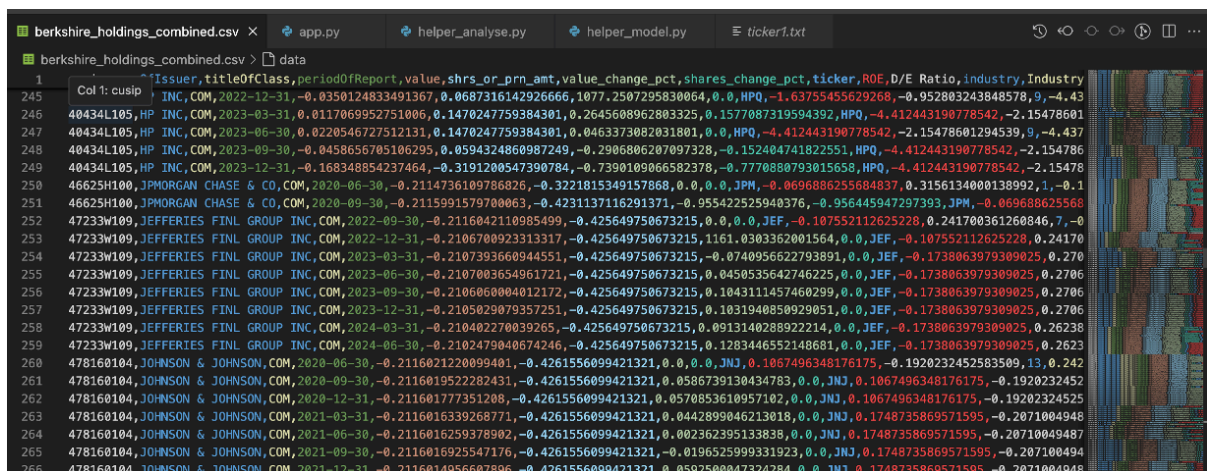
Our research hence seeks to incorporate both the quantitative and qualitative features and make an effort to replicate institutional trading patterns, utilising the public filings SEC Form 13F, with a focus on the successful institutional trading company Berkshire Hathaway. ML will help replicate Berkshire's concentrated and well-curated portfolio, mimicking institutional investors' behaviours by incorporating both price and fundamental data like ROE, and D/E ratios, into ML models and potentially optimise with reinforcement learning methods. Firstly, by adding relevant features from fundamental analysis to the ML models through feature engineering,

we aim to build a classification or regression model to determine stock choices based on historical filings. Beyond portfolio construction, we further incorporate the qualitative and quantitative data into ML models like Long Short-Term Memory (LSTM), to capture time-series patterns, and RF for more feature engineering, aiming to build a prediction model for buy/sell signals. Lastly, there may be a potential utilisation of reinforcement learning for performance optimisation.

Methodology

1. Data initialisation, data scraping

Overall the whole technical component of this project is done in Python language. The initial dataset is obtained directly from the Securities and Exchange Commission's (SEC) Form 13F primarily through web scraping. Given the large filesystem, we fetch web pages and target files with Python's Requests library. BeautifulSoup is utilised for parsing HTML pages for links and navigating through the enormous directories to look for the major information table in XML format as we required. ElementTree is used later for parsing the XML files to retrieve data columns, which include basic information like the name and class, the CUSIP number, value and the number of shares. The Pandas library is then utilised to structure these data into a CSV file to improve the readability of data obtained and ready for further data preprocessing and feature engineering.



	Issuer	titleOfClass	periodOfReport	value	shrs_or_prn_amt	value_change_pct	shares_change_pct	ticker	ROE	D/E Ratio	industry	Industry
245	Col 1: cusip	INC, COM	2022-12-31	0.0350124833491367	0.0687316142926666	1077.2507295830064	0.0	HPQ	-1.63755455629268	-0.952803243848578	9	-4.43
246	40434L105, HP	INC, COM	2023-03-31	0.01170699527512131	0.1470247759384301	0.2645608962803325	0.1577087319594392	HPQ	-4.412443190778542	-2.15478601		
247	40434L105, HP	INC, COM	2023-06-30	0.0220546727512131	0.1470247759384301	0.0463373082031801	0.0	HPQ	-4.412443190778542	-2.15478601	294539	9
248	40434L105, HP	INC, COM	2023-09-30	0.0458656705106295	0.0594324860987249	-0.2906806207097328	-0.152404741822551	HPQ	-4.412443190778542	-2.154786		
249	40434L105, HP	INC, COM	2023-12-31	-0.168348854237464	-0.3191200547390784	-0.7390109066582378	-0.7770880793015658	HPQ	-4.412443190778542	-2.15478		
250	46625H100, JPMORGAN CHASE & CO	COM	2020-06-30	-0.2114736109786826	-0.3221815349157868	0.0	0.0	JPM	-0.0696886255684837	0.3156134000138992	1	-0.1
251	46625H100, JPMORGAN CHASE & CO	COM	2020-09-30	-0.2115991579700063	-0.4231137116291371	-0.955422525940376	-0.956445947297393	JPM	-0.069688625568			
252	47233W109, JEFFERIES FINL GROUP	INC, COM	2022-09-30	-0.2116042110985499	-0.425649750673215	0.0	0.0	JEF	-0.107552112625228	0.241700361260846	7	-0
253	47233W109, JEFFERIES FINL GROUP	INC, COM	2022-12-31	-0.2106700923313317	-0.425649750673215	1161.0303362001564	0.0	JEF	-0.107552112625228	0.24170		
254	47233W109, JEFFERIES FINL GROUP	INC, COM	2023-03-31	-0.2107393660944551	-0.425649750673215	-0.0740956622793891	0.0	JEF	-0.1738063979309025	0.270		
255	47233W109, JEFFERIES FINL GROUP	INC, COM	2023-06-30	-0.2107003654961721	-0.425649750673215	0.0450535642746225	0.0	JEF	-0.1738063979309025	0.2706		
256	47233W109, JEFFERIES FINL GROUP	INC, COM	2023-09-30	-0.2106060004012172	-0.425649750673215	0.1043111457460299	0.0	JEF	-0.1738063979309025	0.2706		
257	47233W109, JEFFERIES FINL GROUP	INC, COM	2023-12-31	-0.2105023979357251	-0.425649750673215	0.1031940050929051	0.0	JEF	-0.1738063979309025	0.2706		
258	47233W109, JEFFERIES FINL GROUP	INC, COM	2024-03-31	-0.210402278039265	-0.425649750673215	0.0913140288922214	0.0	JEF	-0.1738063979309025	0.26238		
259	47233W109, JEFFERIES FINL GROUP	INC, COM	2024-06-30	-0.2102479040674246	-0.425649750673215	0.1283446552148681	0.0	JEF	-0.1738063979309025	0.2623		
260	478160104, JOHNSON & JOHNSON	COM	2020-06-30	-0.2116021220099401	-0.4261556099421321	0.0	0.0	JNJ	0.1067496348176175	-0.1920232452583509	13	0.242
261	478160104, JOHNSON & JOHNSON	COM	2020-09-30	-0.2116019522282431	-0.4261556099421321	0.0586739130434783	0.0	JNJ	0.1067496348176175	-0.1920232452		
262	478160104, JOHNSON & JOHNSON	COM	2020-12-31	-0.211601777351208	-0.4261556099421321	0.0570853610957102	0.0	JNJ	0.1067496348176175	-0.19202324525		
263	478160104, JOHNSON & JOHNSON	COM	2021-03-31	-0.2116016339268771	-0.4261556099421321	0.0442899046213018	0.0	JNJ	0.1748735869571595	-0.2071004948		
264	478160104, JOHNSON & JOHNSON	COM	2021-06-30	-0.2116016259378902	-0.4261556099421321	0.002362395133838	0.0	JNJ	0.1748735869571595	-0.2071004948		
265	478160104, JOHNSON & JOHNSON	COM	2021-09-30	-0.2116016925547176	-0.4261556099421321	0.0196525999331923	0.0	JNJ	0.1748735869571595	-0.207100494		
266	478160104, JOHNSON & JOHNSON	COM	2021-12-31	-0.2116014956607896	-0.4261556099421321	0.0592500047324784	0.0	JNJ	0.1748735869571595	-0.2071004948		

2. Preprocessing

The previously mentioned columns are certainly insufficient to give meaningful insight into our model; hence, we opted for feature engineering in various aspects. We have added multiple columns, including percentage change in value and shares, Return on equity (ROE), debt-to-equity (D/E) ratio, etc. These data require calculation from features not available immediately from our scraped data, which we decided to seek for the yfinance API, which provides stock details by their ticker. The

mapping of tickers to CUSIPs in our data is assisted by combining 3 SEC summary files for initial mapping, openfigi API for the unmatched ones, as well as a manual Python library for updated tickers due to inconsistent representations on Yahoo Finance and SEC filings, e.g. LENB and LEN-B, and in cases where certain tickers are no longer available. With the data obtained from yFinance, we include additional columns that potentially be beneficial to indicate stock performances like profit margin for profitability measurements, industry, Price-to-Book (P/B) Ratio, Price-to-Earnings (P/E) Ratio to identify stock value, where we opted for trailing earnings per share (EPS) instead of forward EPS as it reflects past data so that it is more aligned with Berkshire's long-term approach. We have also included beta to indicate stability, which is preferred in Berkshire's long-term approach, as well as a weighted dividend score based on the stock's dividend policy, with a 50% payout ratio, 30% dividend growth and 20% dividend yield in the calculations, as Berkshire's approach tend to value low but sustainable payout ratio, long term dividend growth, with immediate income represented by dividend yield less valued, with the emphasis in sustainability and growth over high-yield immediate income. We have also included a weighted recommendation score generated by the yFinance recommendation summary information that weights the most recent average score the most. In terms of data cleaning, we have filled missing values with methods like filling default values, forward/backward fill, removing rows, etc. We have also handled data outliers like skewed or non-normal data with the IQR method and encoded categorical data with the label encoding method before scaling the data through standardisation which handles diverse ranges better than methods like normalisation.

3. Machine Learning Approach

As discussed before, a classification or regression model will be utilised with the classification approach later rejected as we opted for only Berkshire's data. After careful consideration, we developed a ranking model, aiming to determine relative importance or priority that emulates Berkshire's prioritisation in its current models. We thus added a new target column for the machine learning (ML) model based on the descending order of value of stocks by period of report. Lightgbm model with LambdaMART algorithm and a custom split for training and testing sets is chosen

and is evaluated by the normalised discounted cumulative gain (NDCG) score, with a value closer to 1 indicating better alignment between predicted and actual rankings.

Results

For quantitative results, the NDCG score of our final ML model with custom split obtained a score of 0.89, with feature importance rank from beta, D/E ratio, P/B ratio, P/E ratio, profit margin, and dividend score. Based on the rankings, we have also provided a final print of top 10 stock recommendations based on ranking more weighted towards recent periods per CUSIP.

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
Early stopping, best iteration is:
[8] training's ndcg@3: 0.935695 training's ndcg@5: 0.93917 training's ndcg@10: 0.946418 valid_1's ndcg@3: 0.897992 valid_1's ndcg@5: 0.916192 valid_1's ndcg@10: 0.939771
NDCG@5: 0.8909430195129692
Feature Importance
3 beta 63.447197
0 D/E Ratio 55.410821
2 P/B ratio 47.814462
4 P/E ratio 38.507655
1 profit_margin 22.805083
5 dividend_score 20.195150
```

```
Final Stock Recommendations (Top 10 Ranked):
CUSIP: 037833100, Issuer: APPLE INC, Weighted Final Rank: 0.00
CUSIP: 060505104, Issuer: BANK AMER CORP, Weighted Final Rank: 0.96
CUSIP: 025816109, Issuer: AMERICAN EXPRESS CO, Weighted Final Rank: 2.39
CUSIP: 191216100, Issuer: COCA COLA CO, Weighted Final Rank: 2.86
CUSIP: 166764100, Issuer: CHEVRON CORP NEW, Weighted Final Rank: 4.72
CUSIP: 674599105, Issuer: OCCIDENTAL PETE CORP, Weighted Final Rank: 5.09
CUSIP: 500754106, Issuer: KRAFT HEINZ CO, Weighted Final Rank: 5.36
CUSIP: 615369105, Issuer: MOODYS CORP, Weighted Final Rank: 6.19
CUSIP: 23918K108, Issuer: DAVITA HEALTHCARE PARTNERS I, Weighted Final Rank: 8.20
CUSIP: 23918K108, Issuer: DAVITA INC, Weighted Final Rank: 8.20
```

Discussion

- The overall technical part of the project was smooth with some challenges. Firstly, the relevant stock data obtained from SEC filings are identified by CUSIPs whereas yFinance API requires a ticker. It was complicated to get all mapping of the ticker to CUSIP at once from SEC files or APIs solely, hence we ended up combining 3 SEC files and used openfigi API for mapping. Furthermore, the ticker identifiers mapped vary with those to be identified by yFinance. As a result, a manual Python dictionary was created to update relevant tickers. With data obtained mostly within the range of the year 2020 to 2024, some companies in the dataset are no longer public due to acquisition and are not appropriate to be included in the dataset anymore, such as Activation Blizzard Inc. which was acquired by Microsoft in January 2023 and ticker AVTI is no longer identifiable by yFinance API. Manual

mapping of tickers can also be challenging due to various stock types - we have attempted to manually map AVTI to AIY.DE which has the same company name after manual research but later realised it was a mutual fund held by the same company and is not a common stock that we were looking for, hence a careful manual mapping is necessary for this project's success. There were also entries like VOO and SPY which are Exchange-Traded Fund (ETF) that tracks a basket of stocks instead of monitoring a single company and thus also not applicable in our data and explains why further data scraping from yFinance API was irrelevant. It is also noted that yFinance API mostly does not provide data in 2024, the year range currently existing in our data set, hence explaining why several columns derived from yFinance API in some year ranges are left empty unwantedly ~~initially~~. For these entries which may still be considered applicable in our project, we have filled missing columns in methods discussed previously.

In the process of feature engineering, we have also considered efficiency issues where every new column in a helper function incurs a new fetch of yFinance API which is inefficient given the request limits. We have thus attempted to save initial fetched data for each ticker in the initial data set in JSON formats for their information, and CSV formats for their data in the balance sheets, financials and cashflows, so that future feature engineering can fetch local data directly without encountering request limits that hinder efficiency of labour. However, the nature of data in which time series data often exists in two-dimensional formats inflicts complications that lead to decisions to abandon this attempted implementation of enhancement. In the dividend score column, our current approach focuses on a uniform weighted ratio of 50% payout ratio, 30% dividend growth and 20% dividend yield to emphasise sustainability and growth over high-yield immediate income aligning with our major focus in Berkshire's portfolio that values long-term growth of capital. In future work, we can potentially implement a more customised payout ratio score adjusted by the corresponding industry with more mature industries allowing a higher threshold.

In terms of the machine learning model, we have initially included all quantitative features which include beta, P/E ratio, ROE, P/B ratio, profit margin, dividend score,

D/E ratio, value percentage change, and shares percentage change. We utilised the default 'train test split' method to split datasets, and results showing value and shares percentage change has little to no importance with the overall NDCG score being 0.936 and the model stopped early at the second iteration.

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
Early stopping, best iteration is:
[2]   training's ndcg@3: 0.917892   training's ndcg@5: 0.917815   training's ndcg@10: 0.929138   valid_1's ndcg@3: 0.949359   valid_1's ndcg@5: 0.945725   valid_1's ndcg@10: 0.971442
NDCG@5: 0.9363916160371878
  Feature Importance
4      beta      31.855407
5      P/E ratio  27.306086
0          ROE    15.807610
3      P/B ratio  11.398039
2  profit_margin  10.249613
6  dividend_score  5.762802
1      D/E Ratio   2.199303
7  value_change_pct  0.160049
8  shares_change_pct  0.000000
```

These features were hence removed in further tuning, with the result significantly improved to a NDCG score of 0.974 and the model was trained for more iteration before stopping at iteration 19, indicating better stability and learning.

```
[LightGBM] [warning] No further splits with positive gain, best gain: -inf
Early stopping, best iteration is:
[19]  training's ndcg@3: 0.950508   training's ndcg@5: 0.955555   training's ndcg@10: 0.963604   valid_1's ndcg@3: 0.935781   valid_1's ndcg@5: 0.95882   valid_1's ndcg@10: 0.968364
NDCG@5: 0.9744438079461284
  Feature Importance
3      P/B ratio  151.886753
1      D/E Ratio  105.086664
5      P/E ratio  100.277453
4          beta   76.231921
0          ROE    38.917024
2  profit_margin  23.887222
6  dividend_score  23.188664
```

Despite having satisfying improvement in scores, we have further evaluated the splitting method utilised in the ML model. The current default splitting method randomly splits data, whereas our project intends to study stock prediction involving time series data. Thus, instead of random splitting, we have introduced a custom split that takes a portion of data entries per CUSIP ordered by period of report, with earlier entries per CUSIP in training and later entries in testing that is more realistic than a random split. Despite the final model after further feature engineering results in a slightly lower NDCG score of 0.89, it is necessary to note that this does not imply the new method is unsuitable with worse performance, but more likely indicating that the previous random approach may be overfitting and overestimating real-world performance with data distribution changes over time.


```

[LightGBM] [warning] NO further splits with positive gain, best gain: -inf
Early stopping, best iteration is:
[8]    training's ndcg@3: 0.935695    training's ndcg@5: 0.93917    training's ndcg@10: 0.946418    valid_1's ndcg@3: 0.897992    valid_1's ndcg@5: 0.916192    valid_1's ndcg@10: 0.939771
NDCG@5: 0.8909430195129692
Feature Importance
3      beta      63.447197
0      D/E Ratio  55.410821
2      P/B ratio  47.814462
4      P/E ratio  38.507655
1      profit_margin  22.805083
5      dividend_score  20.195150

```

In terms of further evaluation, we can potentially further incorporate more qualitative factors and evaluate our model with the market index or benchmark such as Dow Jones Industrial Averages. In our improvement attempt, we tried to incorporate the recommendation summary information from yFinance API by the extra recommendation score column as mentioned previously. However the attempt was reverted, as after careful consideration it is more relevant to short-term grading and evaluation given the recommendation data obtained from the API is within the recent 3-4 months, hence irrelevant to data scraped from SEC filings which are mostly 2023 or prior, and does not align to replicate Berkshire's portfolio in long term growth capital. While for extra referencing purposes, we have not included the recommendation feature in our model but decided to print rankings of both our weighted recommendation score and recommendation mean scraped directly from yFinance API to expand perspectives in considerations. It is observed that both rankings differ from the one generated by our model in the evaluation of Berkshire's holding values. It can be explained by the fact that Berkshire values more long-term investment while yFinance recommendation summary is based on recent analysts' relatively short-term opinions. It is also noticed that the two rankings derived from the API have slightly different recommendation scores, hence rankings. A possible reason is the calculation approach, our manual recommendation score was calculated by getting the average score per period and weighting them by a period with more weighting in recent periods, while the recommendation means may have a different approach to getting the value, which is not clear currently. But overall, both of these recommendation metrics were derived from the recommendation summary provided by yFinance API, which is also not well represented and valid to be incorporated into the model, and very questionable for its accuracy and application.

	period	strongBuy	buy	hold	sell	strongSell
0	0m	8	24	12	1	2
1	-1m	8	24	12	1	2
2	-2m	8	23	12	1	2
3	-3m	8	24	12	0	2

Recent top 10 recommendation by analysts (Recommendation SCORE):
 CUSIP: G7709Q104, Issuer: ROYALTY PHARMA PLC, Recommendation Score: 0.82
 CUSIP: 023135106, Issuer: AMAZON COM INC, Recommendation Score: 0.80
 CUSIP: 874039100, Issuer: TAIWAN SEMICONDUCTOR MFG LTD, Recommendation Score: 0.79
 CUSIP: 609207105, Issuer: MONDELEZ INTL INC, Recommendation Score: 0.79
 CUSIP: 166764100, Issuer: CHEVRON CORP NEW, Recommendation Score: 0.78
 CUSIP: 881624209, Issuer: TEVA PHARMACEUTICAL INDS LTD, Recommendation Score: 0.78
 CUSIP: 57636Q104, Issuer: MASTERCARD INC, Recommendation Score: 0.76
 CUSIP: 92826C839, Issuer: VISA INC, Recommendation Score: 0.76
 CUSIP: 872590104, Issuer: T-MOBILE US INC, Recommendation Score: 0.75
 CUSIP: 060505104, Issuer: BANK AMER CORP, Recommendation Score: 0.75

Recent top 10 recommendation by analysts (Recommendation MEAN):
 CUSIP: 531229748, Issuer: LIBERTY MEDIA CORP DEL, Recommendation Score: 1.00
 CUSIP: 531229722, Issuer: LIBERTY MEDIA CORP DEL, Recommendation Score: 1.00
 CUSIP: 874039100, Issuer: TAIWAN SEMICONDUCTOR MFG LTD, Recommendation Score: 0.91
 CUSIP: 023135106, Issuer: AMAZON COM INC, Recommendation Score: 0.90
 CUSIP: 609207105, Issuer: MONDELEZ INTL INC, Recommendation Score: 0.89
 CUSIP: 531229854, Issuer: LIBERTY MEDIA CORP DEL, Recommendation Score: 0.88
 CUSIP: G7709Q104, Issuer: ROYALTY PHARMA PLC, Recommendation Score: 0.85
 CUSIP: 58933Y105, Issuer: MERCK & CO. INC, Recommendation Score: 0.84
 CUSIP: 58155Q103, Issuer: MCKESSON CORP, Recommendation Score: 0.84
 CUSIP: 57636Q104, Issuer: MASTERCARD INC, Recommendation Score: 0.83

Conclusion

This research presents a ML model to replicate an institutional investment portfolio specifically by the renowned Berkshire's holdings, which is further validated by feature rankings and high NDCG scores. Future work could explore incorporating more qualitative factors, and more robust data engineering and ML frameworks for further improvements.

Reference

1. Team, T.I. (2024) *Warren Buffett's investment strategy*, Investopedia. Available at: <https://www.investopedia.com/articles/01/071801.asp> (Accessed: 03 October 2024).
2. McFarlane, G. (2024) *How Warren Buffett made Berkshire Hathaway a winner*, Investopedia. Available at: <https://www.investopedia.com/articles/markets/041714/how-warren-buffett-made-berkshire-hathaway-worldbeater.asp> (Accessed: 03 October 2024).
3. *Stock market prediction* (2024) Wikipedia. Available at: https://en.wikipedia.org/wiki/Stock_market_prediction#:~:text=Alongside%20the%20patterns%2C%20techniques%20are,widely%20used%20by%20technical%20analysts. (Accessed: 03 October 2024).
4. Martin, G.S. and Puthenpurackal, J. (2005) *Imitation is the sincerest form of flattery: Warren Buffett and Berkshire Hathaway*, SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=806246 (Accessed: 06 October 2024).
5. Snow, D., 2019. Machine learning in asset management—Part 1: Portfolio construction—Trading strategies. *The Journal of Financial Data Science*.
6. Ryll, L. and Seidens, S. (2019) *Evaluating the performance of machine learning algorithms in Financial Market Forecasting: A comprehensive survey*, arXiv.org. Available at: <https://arxiv.org/abs/1906.07786> (Accessed: 06 October 2024).
7. Jiao, Y. and Jakubowicz, J., 2017, December. Predicting stock movement direction with machine learning: An extensive study on S&P 500 stocks. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 4705-4713). IEEE.
8. Griffin, J.M., Harris, J.H. and Topaloglu, S., 2003. The dynamics of institutional and individual trading. *The Journal of Finance*, 58(6), pp.2285-2320.