Development of Model for Stock Performance Prediction

MS 499 Project Course

Hrriday V. Ruparel, 22110099
B.Tech. Computer Science and Engineering
Indian Institute of Technology, Gandhinagar

April 27, 2024

Under the guidance of
Prof. Kaisivisvanathan Chelvakumar
IIT Gandhinagar

1 Abstract

This project delves into developing a deep learning-based estimator for US S&P 500 stock performance prediction, integrating financial, technical, and macro-economic factors, through a case study on the highly acclaimed US company, Berkshire Hathaway (BRK). A byproduct of this research is a `Stock Analysis Dashboard v2.0` offering comprehensive financial insights into S&P 500 stocks and inferencing capabilities.

2 Introduction

Stock performance prediction is a critical aspect of financial analysis, essential for making informed investment decisions. However, accurately forecasting stock prices remains challenging due to the complex, non-linear and volatile nature of financial markets. Historically, a variety of quantitative and qualitative approaches have been utilized for stock price prediction, including fundamental and technical analysis.

In recent years, machine learning algorithms have gained popularity for their ability to detect patterns and correlations among various factors affecting a variable of interest from large datasets and unlike classical statistical models, these algorithms are capable of forecasting time series data without much assumptions about the data.

Among the vast ocean of techniques available for stock price predictions, this research attempts to achieve reliable predictions despite market unpredictability, by combining important factors affecting the performance of stock prices and utilizing deep learning techniques to achieve the goal. I, thus propose, `KTEstimator`, a simple Multi-Layer Perceptron (MLP) based model capable of forecasting stock prices[1].

As a virtue of the research, a `Stock Analysis Dashboard v2.0` was also developed that provides a one-stop destination to perform financial analysis of desired S&P 500 stock, visualizing core financials of companies, analyzing technical aspects and inferring from comparisons with benchmarks and competitors.

3 Research & Intuition

The project course started with the introduction of financial statements of companies like balance sheets and income statements. Much inspired from the initial interactions, I came to believe that understanding the importance of these powerful tools that provide an in-depth insight into the internal functioning of companies is quite essential as these can be used as internal factors in modelling the performance of a company's shares. However, a company's internals are a direct or an indirect consequence of the external market forces and macro-economic influences driving the company's decisions. Realizing this is crucial for building intuition behind the model developed as the features/factors that the model utilizes to make stock price predictions are effectively the internals of the company (like assets, liabilities, revenues, income, etc.) and external market influences (like recession, country's GDP, inflation, risk-free rates, Gold reserves, etc.).

To validate the above claims and intuition, it was necessary to analyze the reactions of stock prices to changes in the internal and external factors of companies. Hence, as result of this analysis, a comprehensive `Stock Analysis Dashboard v2.0` was built using `Python` programming language and a well-known Python library for financial and technical analysis, `yfinance`. The Dashboard enables users to enter a US S&P 500 Stock symbol and produces a detailed financial report of the company comprising of the following sections:

1. Short summary of the company.

[1] To a certain degree of unpredictability and error

2. Latest news related to the company.

3. Investor Data.

4. Selected latest Financials and Financial Ratios of the company.

5. Technical Analysis section comprising of candlestick chart of the company's historical stock data along with historical US macro-economic factors like US Inflation Rate[2] and US GDP-based Recession Indicator[3].

6. Competitor Analysis and Benchmarked Returns chart offering comparison of performance of the company's share against competitors and well-known benchmarks (like S&P 500 index and corresponding sub-industry index).

Without loss of generality, a case-study was conducted on Berkshire Hathaway Inc. (BRK-B), a financial company known to constantly beat the market in terms of returns. The following images (Figure 1 - Figure 4) were taken from the `Stock Analysis Dashboard v2.0` showcasing the analysis for Berkshire Hathaway Inc. (BRK-B).

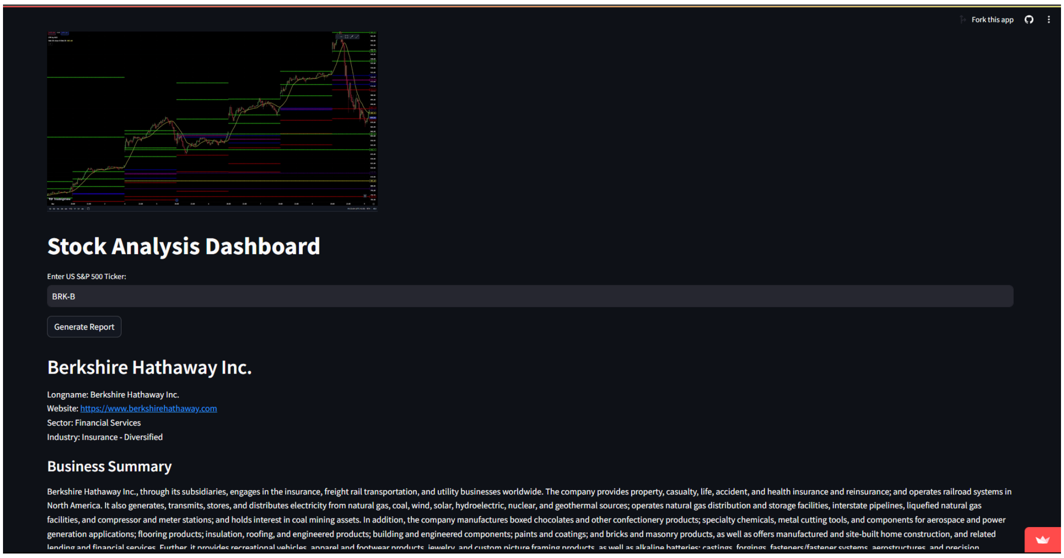

Figure 1: `Stock Analysis Dashboard v2.0` for BRK-B

[2] Acquired from Federal Reserve Economic Data (FRED)
[3] Acquired from Federal Reserve Economic Data (FRED)

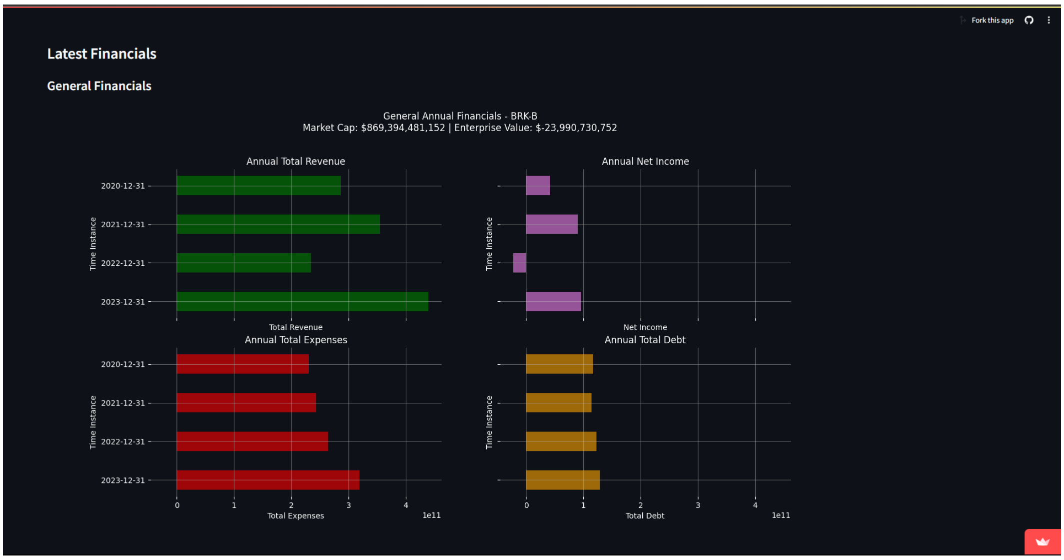

Figure 2: Latest Financials for BRK-B

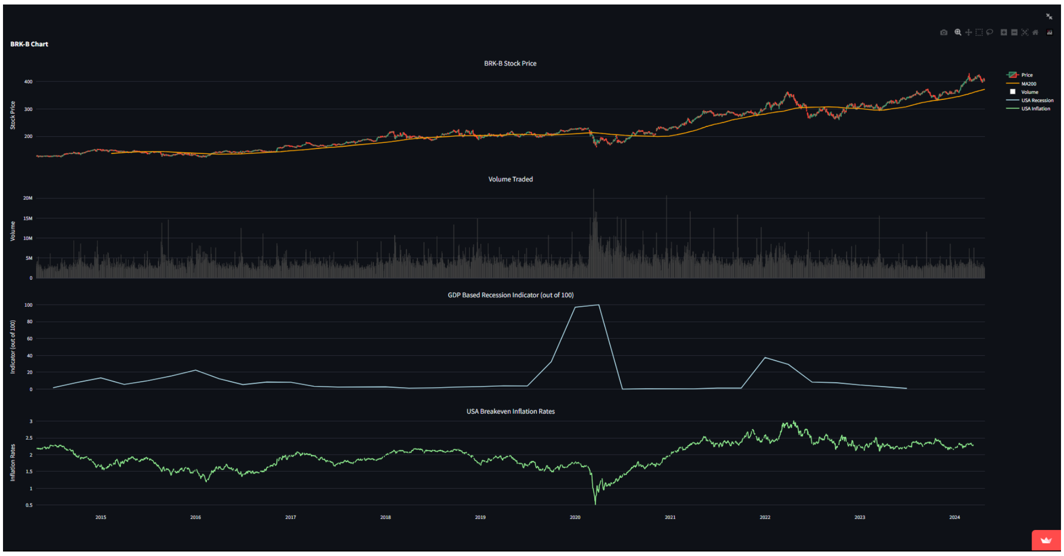

Figure 3: Technical Analysis for BRK-B

Figure 4: Benchmarked Returns over the last decade for BRK-B

Figure 5: Stock Price Analysis with respect to internal and external factors for BRK-B

BERKSHIRE HATHAWAY INC.
and Subsidiaries
CONSOLIDATED STATEMENTS OF EARNINGS
(dollars in millions except per share amounts)
(Unaudited)

Investment and derivatives making loss!

	First Quarter	
	2022	**2021**
Revenues:		
Insurance and Other:		
Insurance premiums earned	$ 17,492	$ 16,424
Sales and service revenues	37,862	33,698
Leasing revenues	1,672	1,324
Interest, dividend and other investment income	1,862	1,851
	58,888	53,297
Railroad, Utilities and Energy:		
Freight rail transportation revenues	5,944	5,378
Energy operating revenues	4,818	4,848
Service revenues and other income	1,160	1,076
	11,922	11,302
Total revenues	70,810	64,599
Investment and derivative contract gains (losses)	(1,978)	5,700

Figure 6: Internal Factors - Income Statement - Investments and Derivatives

BERKSHIRE HATHAWAY INC.
and Subsidiaries
CONSOLIDATED STATEMENTS OF COMPREHENSIVE INCOME
(dollars in millions)
(Unaudited)

Net Earnings Declined

	First Quarter	
	2022	**2021**
Net earnings	$ 5,585	$ 11,840
Other comprehensive income:		
Unrealized appreciation of investments	(236)	(87)
Applicable income taxes	51	20
Foreign currency translation	(316)	(285)
Applicable income taxes	(11)	(3)
Defined benefit pension plans	26	61
Applicable income taxes	(5)	(20)
Other, net	87	(6)
Other comprehensive income, net	(404)	(320)
Comprehensive income	5,181	11,520
Comprehensive income attributable to noncontrolling interests	122	136
Comprehensive income attributable to Berkshire Hathaway shareholders	$ 5,059	$ 11,384

Figure 7: Internal Factors - Income Statement - Net Earnings

Certain inferences have been drawn (Figure 5 - Figure 7):

1. Whenever fresh quarterly reports of BRK-B were released (usually the month-ends of March, June, September, December), a price reaction was noticed depending on the internals of the company; a decline in earnings or income was reflected by a bearish movement while a positive outlook pushed the stock price higher. This is quite natural and intuitive but the main observation is the major reaction of the stock price around the company's quarterly financial filings.

2. The price trend thus set by the price reaction generally continued throughout the consecutive quarter, but there were contrary cases as well. These instances could be attributed to external market influences like Covid-19 Crisis and extreme inflation.

Thus, on the basis of these observations, it was concluded to utilize the internal and external factors mentioned as before in modelling the stock performance of Berkshire Hathaway Inc.

4 Prediction Model

In finance, especially in predicting stock prices, deep learning methods like Multi-Layer Perceptrons (MLPs) have become increasingly popular because of their over-powerful characteristic abilities to learn any "input-ouput mapping", provided the inputs and desired outputs. Think of MLPs as "black-boxes" that can tune itself to fit to any data like a statistical measure and base predictions on the "learnt" input-output mappings.

Another analogy is that MLPs are like super-smart students who learn from examples. Imagine a math problem, and a teacher wants to teach the student how to solve it. The teacher (here, us) would give the student (here, the model) lots of examples (here, collected financial times series), and over time, the student would start to see patterns in how the problems are solved. That's what an MLP does with financial data. It looks at a lot of past data, like quarterly financial reports, and tries to find patterns in it. Then, when you give it new data, it uses those patterns to make predictions. Training an MLP is like teaching that student how to solve math problems. You give them more and more examples, correct their methods according to best outcomes, and eventually, they get really good at solving similar problems on their own.

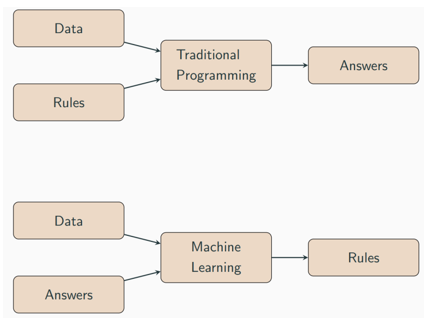

Figure 8: Traditional Programming vs Deep/Machine Learning

Earlier, algorithms were designed in such a manner that the programmer used to set down the "rules": given an input, follow the programmed "rules" to get to the output. With deep learning techniques, we feed in the inputs and desired outputs, and the model learns the "rules" on its own. Once it has learned these rules, the model then is capable to give outputs based on "learned rules" on new unseen inputs.

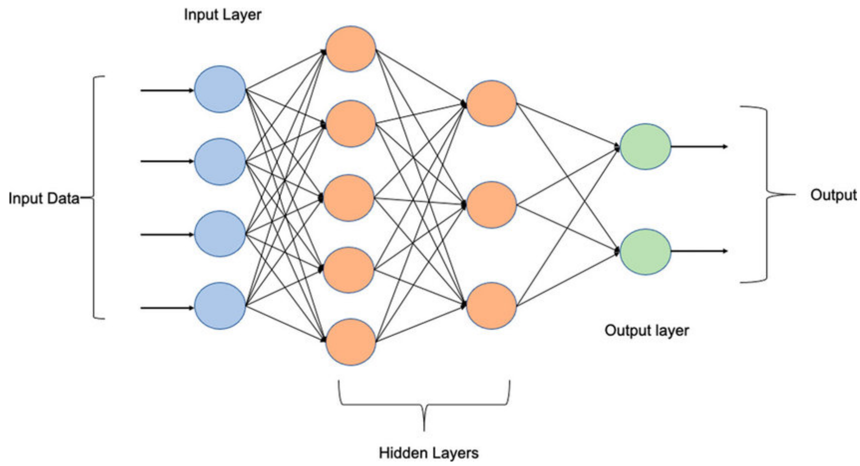

Figure 9: A standard Multi-Layer Perceptron (MLP)

Therefore, the `KTEstimator` model is a Multilayer Perceptron (MLP) to predict stock prices based on past financial data (both internal and external factors included). An MLP is like a stack of interconnected layers, each layer consisting of nodes. The MLP in this case has an input layer where past financial data for a certain number of quarters (K quarters) is fed in. These inputs are then passed through two hidden layers, each containing multiple nodes. These hidden layers act as filters and feature extractors, processing the inputs and extracting important features that help in predicting the stock price for the next quarter-end (T = 1).

Each node in the hidden layers takes a weighted sum of the inputs from the previous layer and applies an activation function to produce an output. These outputs are then passed to the next layer, where the process repeats. Finally, the MLP has an output layer where the predicted stock price for the next quarter is generated based on the processed information from the hidden layers.

Through a process called backpropagation, the MLP adjusts the weights (parameters) of its connections between nodes to minimize the difference between its predicted stock prices and the actual stock prices in the training data (historical data separated out for calibration for model). This way, the MLP learns to capture complex patterns and relationships in the financial data, enabling it to make predictions.

5 Data Processing for Prediction Model

The following points would help understand the representation of financial data as depicted in Figure 10.

- Suppose we have a time series of financial data labelled by timestamps a_i, where i denotes the day on which Berkshire releases its quarterly annual reports.
- Each timestamp a_i would be characterized by a financial feature vector $\vec{f_i}$ (depicted in light blue in Figure 10) containing all selected financial factors as per user (both internal and external factors). The feature vector for timestamp a_i would also contain the closing stock price of BRK-A and BRK-B on that day.
- Think of the feature vectors as all the information an observer has till timestamp a_i out of which only the latest K timestamp feature vectors are being used as inputs to the model.
- As time progresses to next financial quarter, the input window slides and we get a new set of K feature vectors. For each set of K input features in our training data

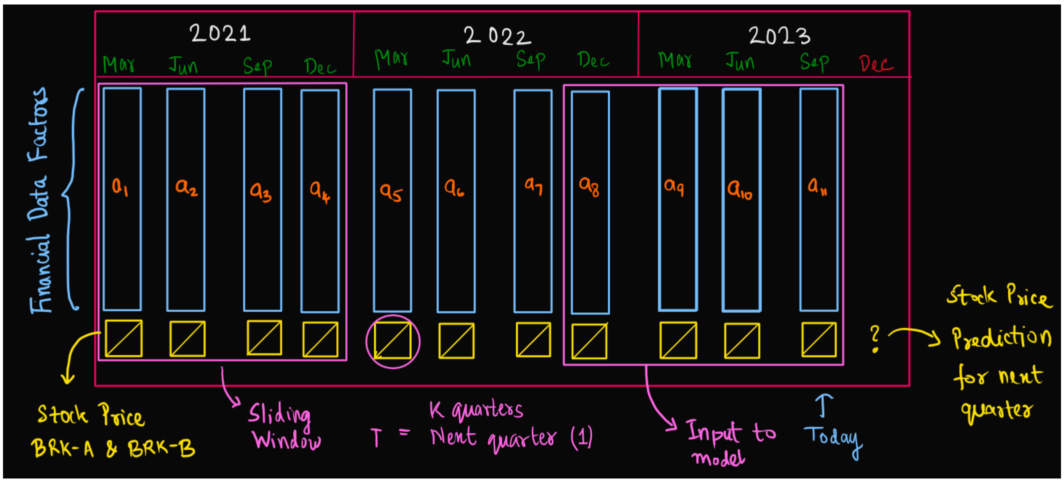

Figure 10: Financial Data Representation for the Model

(historical data filtered out for model calibration purposes) ending at some timestamp, say a_i, we assign the output of the model to be the stock price of BRK-A and BRK-B at timestamp a_{i+1}, call it $\overrightarrow{s_{i+1}}$.

- Hence, the final data representation to be fed into the `KTEstimator` prediction model as "input-output" mapping and for calibration of the model parameters is shown in Figure 11.

For the purpose of implementation of the model, the following variables have been fixed:

- Time duration = 5 years (Jan 2019 - Dec 2023)
- Number of timestamps a_i, L = 20 (= 4 quarters per year x 5 years)
- Number of features per timestamp = 40 (shown in Figure 12)

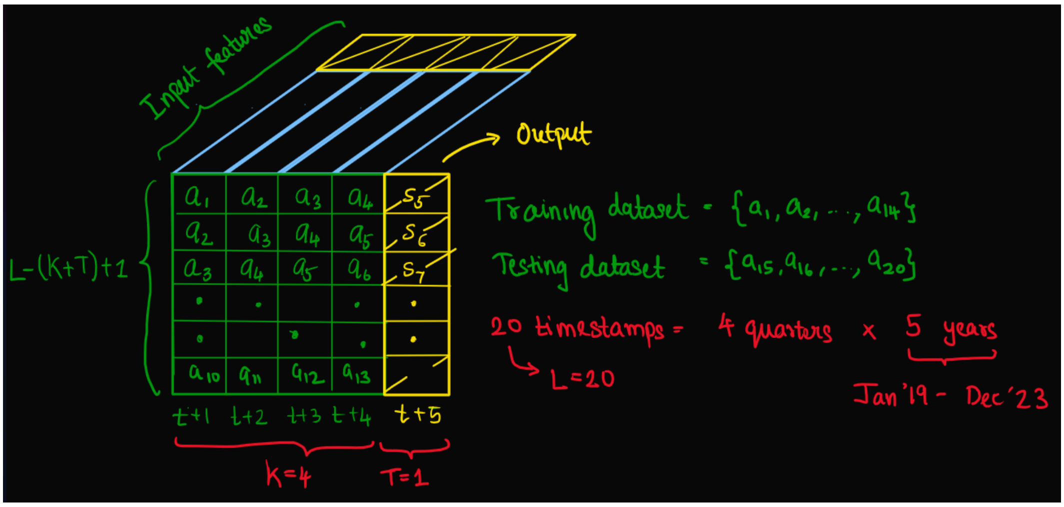

Figure 11: Final Data Representation for the Model - Input Output Mapping

Year	2023				2022				2021				2020				2019			
Month	December 31	September 30	June 30	March 31	December 31	September 30	June 30	March 31	December 31	September 30	June 30	March 31	December 31	September 30	June 30	March 31	December 31	September 30	June 30	March 31
Month No.	12	9	6	3	12	9	6	3	12	9	6	3	12	9	6	3	12	9	6	3
ASSETS																				
Insurance																				
Short-term investment in U.S. Treasury Bills	129619	126401	97322	103869	92774	76332	74803	67145	58535	79209	101760	85385	63822	118906	110518	94623	81506	53378	77745	88029
Fixed Maturity Securities	23758	22435	22353	22566	25128	18602	21136	21718	16434	18125	20460	20027	18685	19435	19210	17923	19898	19172	19962	19415
Equity Securities	353842	318621	353409	328161	308793	306167	327662	390538	350719	310739	307942	282097	248027	245317	207454	180782	172757	220051	200516	191771
Equity Method Investments	29066	27496	27493	26403	28050	28714	17494	17596	17375	16658	16542	16533	17,505	17152	17132	17583	17325	17535	17208	17308
Loans Receiveables	24681	24009	23530	23144	23208	22094	21877	21265	20751	20397	19900	19449	17,527	18584	18175	17660	16280	17135	16807	16432
Property, Plant and Equipment	199646	195214	193160	190181	181381	177850	177288	176722	176364	174123	173652	172789	159276	161539	159902	158522	152408	156556	154719	146599
Equipment held for lease/ Service model	16947	16284	16028	15674	15584	15139	15125	15038	14918	14752	14659	14567	15065	14714	14887	14899	14298	14934	14741	14429
Goodwill	84626	85652	85853	83502	78119	73395	73618	73822	73875	73770	73758	73695	81882	71865	71753	81659	81025	81228	81269	81220
LIABILITIES																				
Insurance and Others																				
Unpaid Losses	145729	143743	143451	143020	142887	128308	125794	125369	124920	125475	123195	121646	115460	119827	117506	115631	110292	113361	112330	111168
Unearned Premiums	30507	31914	31173	30359	28657	26878	25727	25368	23512	25239	23732	23756	19782	22018	20992	21718	18093	20764	20113	20172
Life, Annuity and Insuarance Benefits	20213	18556	19635	19937	22421	22305	22562	22673	22452	22409	22106	21709	20155	20975	20642	20193	18632	19422	19155	18918
Notes payable and other borrowings	128271	124781	125347	123624	122744	116496	119081	119661	114262	114965	115223	114531	103368	107691	106689	104233	97490	102194	89907	97193
Accounts payable	54863	55077	52233	52525	50116	47971	47580	46005	46072	48132	45544	43693	42319	44026	40665	39614	37186	41189	40588	39807
SHAREHOLDER'S EQUITY																				
Capital in excess of par value	34480	34473	35140	35156	35167	35190	35204	35586	35592	35603	35635	35630	35658	35621	35615	35619	35707	35612	35610	35622
Retained earnings	607350	569776	582543	546631	511602	493438	496126	539881	534421	494775	484431	456337	402493	408791	738654	352359	321112	373334	356846	342773
Treasury stock, at cost	-76802	-74655	-73568	-72265	-67826	-64972	-63934	-62906	-59795	-53072	-45446	-39418	-8125	-24075	-14815	-9700	-3109	-5937	-5252	-4799
EXTERNAL FACTORS																				
10-Year Treasury Constant Maturity Rate	3.88	4.59	3.81	3.48	3.88	3.83	2.98	2.32	1.52	1.52	1.45	1.74	0.93	0.69	0.66	0.7	1.92	1.68	2	2.41
10-Year Breakeven Inflation Rate	2.16	2.35	2.22	2.32	2.3	2.15	2.33	2.84	2.56	2.37	2.32	2.37	1.99	1.63	1.34	0.87	1.77	1.53	1.69	1.88
GDP based Recession Indicator	2	2	0.9	3	4.9	7.6	8.3	29.2	37.4	1.2	1.2	0.3	0.2	0.4	0	100	97.1	32.5	3.8	3.9
Unemployment Rate	3.7	3.8	3.5	3.4	3.4	3.6	3.5	3.7	4	4.5	5.4	6.1	6.4	6.8	10.2	14.8	3.6	3.6	3.7	3.7
Equity Market Volatility Tracker: Business Investme	0.53117	0	0.34819	0.15942	1.05738	0.2898	0.79661	0	0	0	0.12134	0.16128	0	0.3439	0.15555	0.35154	0.52741	1.38185	1.36189	1.088
Gold Price	2062.66	1848.46	1919.31	1969.3	1823.58	1660.73	1807.18	1936.41	1828	1756.5	1769.41	1707.52	1898.07	1885.33	1781.93	1574.74	1517.7	1472.2	1409.51	1292.4
INCOME STATEMENT																				
Total Revenue	93466	93120	92503	85393	78165	76934	76180	70810	71907	70583	69114	64599	64450	63024	56840	61265	65368	64972	63598	60678
Investment and derivative contract gains	36814	-29778	33061	34758	14473	-13465	-66919	-1978	40527	4921	27394	5700	-101196	31582	40085	70275	31311	10926	10048	20322
Cost of Sales and Services	30580	31049	30621	30319	31609	31292	31633	29785	29863	28984	28761	26530	23526	25957	26216	25392	27277	26950	27047	25767
Selling, general and administrative expenses	6878	5120	5005	5602	7425	4068	3762	4251	4999	4889	5045	3910	5505	5181	4652	4471	5677	4384	4829	4432
Total Costs and Expense	83764	80653	80640	76087	71232	68027	64974	62396	63592	63397	60934	56022	58725	56966	61259	54339	26479	56053	56011	53521
er class A share	26124	-8824	24775	24337	12349	-1832	-29754	3702	26452	6882	18488	7638	22013	18994	16314	-30653	17892	10119	8608	13209
Net Earnings per class B share	5.62	5.88	16.52	16.25	8.23	-1.22	-19.84	2.47	17.63	4.59	12.33	5.09	14.68	12.66	10.88	-20.44	11.92	6.75	5.74	8.81
Comprehensive income - BH shareholders	38083	-12799	36648	35580	19807	-4747	-45502	5059	40317	9729	28581	11384	36782	30914	27049	-51224	29324	15794	14122	21949
RATIOS																				
P/E Ratio	10.17	8.99	99.46	91.24	90.78	52.87	7.377	8.95	8.06	6.14	6.33	14.26	15.28	23.1	19.6	5.515	6.819	18.84	18.26	18
Diluted EPS	25990	-8824	24775	24377	12274	-1907	-29663	3784	26529	6882	18488	7638	22508	18994	16314	-30653	17884	10119	8608	13209
D/E ratio	0.2285	0.2375	0.2322	0.245	0.2593	0.2558	0.2582	0.2355	0.2257	0.2594	0.2711	0.2805	0.2638	0.2594	0.2711	0.2805	0.2433	0.247	0.2612	0.2635
ROE (%)	17.81	14.79	17.28	1.49	-4.71	-0.21	2.44	16.83	18.63	18.42	23.43	24.12	10.35	8.85	5.55	2.52	20.49	7.1	7.75	7.32
ROA (%)	9.32	7.66	8.92	0.76	-2.44	-0.11	1.29	8.91	9.79	9.58	12.09	12.32	5.23	4.49	2.82	1.28	10.49	3.59	3.93	3.71
ROCE (%) (A)	5.2	5.2	4.7	4.5	4.5	4.9	4.7	4.3	4.3	4.2	4.2	2.7	3.7	3.9	4.4	4.8	4.7	4.9	4.9	5.1
ROIC (%)	4.3	4.2	3.8	5	3.44	0.5	4.8	3.6	3.7	3.5	3.4	2.2	3	3.1	3.4	4.2	3.9	4.1	4.1	4.3
Operating Margin	34.9	29.2	34.42	4.52	-7.62	1.7	7.61	40.06	43.73	43.51	54.58	61.26	30.37	26.65	19.15	7.44	43.04	17.27	18.52	17.45
BRK-A	542625	531477	517810	465600	468711	406470	408950	528921	450662	411379	418601	385702	347815	320001	267300	272000	339590	311832	318350	301215
BRK-B	356.660004	350.299988	341	308.769989	308.899994	267.019989	273.019989	352.910004	299	272.940002	277.920013	255.470001	231.869995	212.940002	178.509995	182.830002	226.5	208.020004	213.169998	200.89

Figure 12: Gathered Financial Data of BRK

6 Results & Inferences

After much parameter fine-tuning, the following results shown in Figure 13 and Figure 14 were obtained for Berkshire Hathaway Inc. for the following best hyperparameter set:

- Number of hidden layers of MLP = 3
- Number of nodes in hidden layers of MLP:
(hidden layer 1, hidden layer 2, hidden layer 3) = (40, 20, 5)
- Percentage of data left for training (calibration) = 70%
- Context length, K = 2
- Forecasting horizon, T = 1

One can see that the model predictions are slightly deviating from the ground truth stock prices of BRK-A and BRK-B, though the model is able to forecast a trend for the future stock prices similar to the ground truth. The model performs moderately well in modelling the relationship of the inputs (financial data) to the outputs (stock prices) with a Normalized Mean Square Error (NMSE) of about $5.342 \text{x} 10^{-3}$ and $9.239 \text{x} 10^{-3}$ for stock price predictions of BRK-A and BRK-B respectively.

$$\text{Normalized Mean Square Error (NMSE)} = \frac{e^T e}{X^T X},$$

where e is error between ground truth, X, and predictions.

The model in essence simulates the ideology of Financial Analysts, who look at historic financial, technical and macro-economic data and propose their estimate of stock price for the next few days/months based on observations from the past experiences.

The following could be potentially the reasons for the model not accurately predicting the stock prices:

1. Stock prices are inherently volatile and subject to sudden changes due to market dynamics, geopolitical events, economic shifts, and other external factors. Predicting stock prices accurately in such a dynamic environment is inherently challenging.
2. MLPs are known to be learners that can approximate input-output mappings upto any degree of approximation, but if not trained well, they might overfit the data leading to loss of generality and poor performance on unseen data.
3. The features used as inputs to the model may not representative of the intrinsic complex factors that actually affect the stock price (one example of this could simply be herd mentality of traders and market speculations, something which cannot be quantified). In such scenarios, the model may not be able to capture the underlying patterns in the data.
4. The scarcity of data for calibration, training and testing can significantly impact the performance of the model. Greater the amount of data, greater is one's expectation of the model's performance. Since in this scenario, the model is trained on ¡ 20 data points (by virtue of the solution proposed), it becomes difficult for the model to learn representations of underlying patterns in the data.

Date	Ground Truth BRK-A	Ground Truth BRK-B	Average Prediction BRK-A	Average Prediction BRK-B	Median Prediction BRK-A	Median Prediction BRK-B
Dec 31 2022	468711	308.899994	439383.06	292.87747	447392.9	297.12042
Mar 31 2023	465600	308.769989	422315.3	280.56223	419685.2	276.78687
Jun 30 2023	517810	341	465373.44	309.37238	483883.53	318.83786
Sep 30 2023	531477	350.299988	496828.75	332.5916	509715.75	340.6574
Dec 31 2023	542625	356.660004	470817.7	312.98398	491054.4	322.52228

Figure 13: Result Table Comparing Mean and Median estimates with unseen Ground Truth stock prices

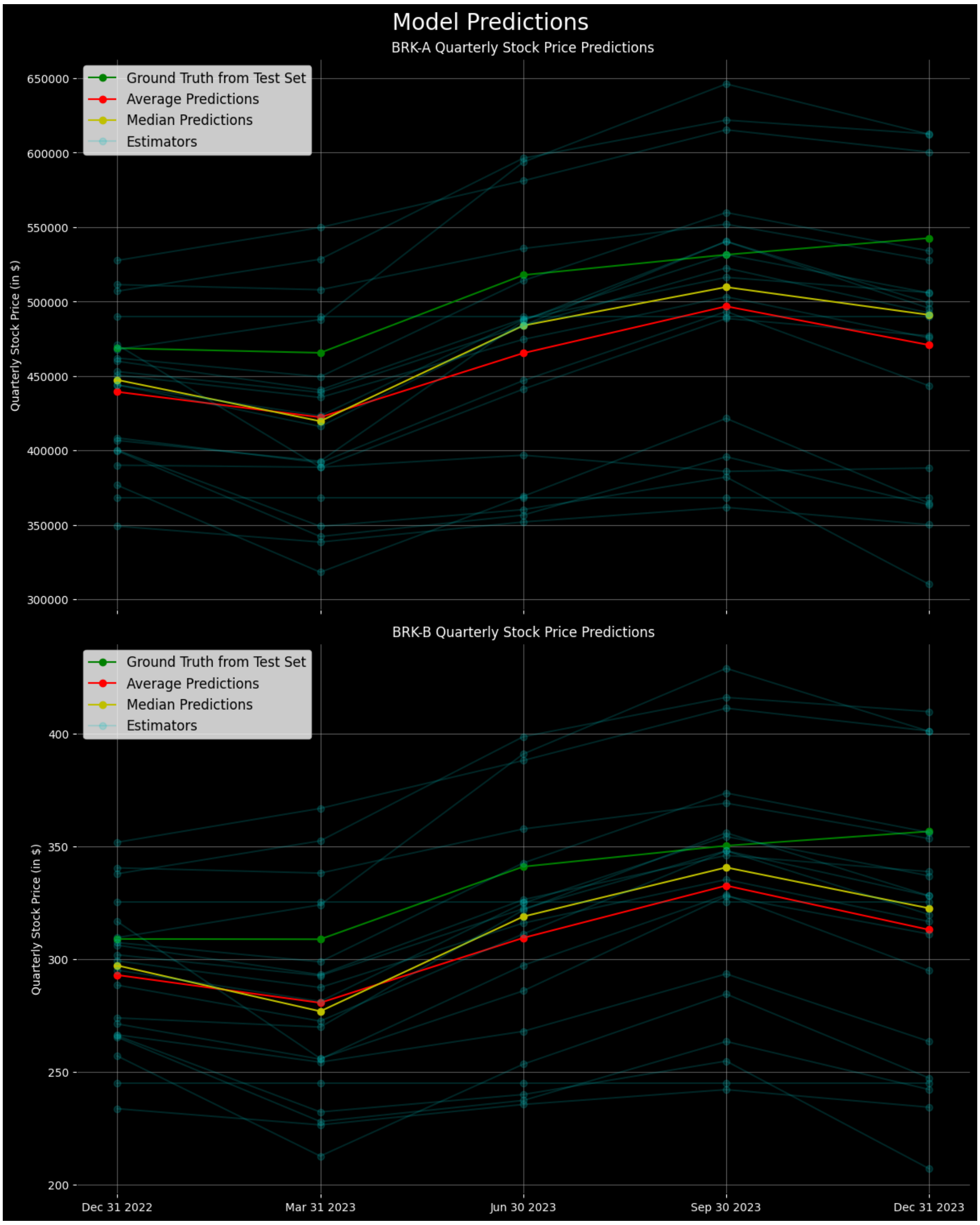

Figure 14: Predictions by the model on unseen data

7 Conclusion & Future Work

The financial research related to stock performance prediction resulted in carefully understanding the dependence of stock price of companies on internal and external economic factors by virtue of which, a `Stock Analysis Dashboard v2.0` was developed to simplify the process of reading financial statements. The Dashboard provides an overall view of the company and researchers can further delve into the fine details after observing some interesting patterns. The purpose of the Dashboard is to facilitate the process of financial analysis and it should not be treated as the only source of financial analysis.

Observations made using the `Stock Analysis Dashboard v2.0` further encouraged to develop a deep learning based stock performance prediction model `KTEstimator` that utilizes the historical financial, technical and macro-economic data to forecast stock prices. For the purpose of implementation, Berkshire Hathaway Inc. (BRK) was chosen as the candidate. Results depicted that the model was able to predict the trend of the stock prices but produced slightly inaccurate forecasts.

As part of future work, the model can be further improved using state-of-the-art techniques like LSTM and Transformers that are better suited for time series predictions. The issue related to scarcity of data is crucial and can be tackled by cleverly modifying the solution proposed, without loss of generality. Better features can be selected and feature transformations can be implemented to model high dimensional data and non-linearity of stock prices.

8 Acknowledgements

I would like to extend my sincere appreciation and gratitude to Professor Chelvakumar for his invaluable guidance and regular insights throughout the course of this research. His expertise and mentorship have greatly contributed to the development and refinement of the research. I am grateful for his unwavering support and constructive feedback, which have played a significant role in shaping the direction of the research.

9 Resources

- Source code of deep learning based stock performance prediction model `KTEstimator` and `Stock Analysis Dashboard v2.0`
- Link to `Stock Analysis Dashboard v2.0`