

STOCK RECOMMENDATION: SITTING ON THE SHOULDER OF GIANTS

**Mini Project-I
Report
Bachelors of Technology
in
Data Science and Artificial Intelligence
by
Ankit Kumar - 20BDS009
Devansh Purwar - 20BDS017
Gurvinder Singh Yadav - 20BDS025
Ripudaman N Singh - 20BDS043**



**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DHARWAD**

JAN-MAY 2023

Abstract

An average person who goes about his life does not have the required skills and aptitude nor does he/she have adequate time to acquire the knowledge required to make sound investment decisions in stocks, derivatives, options, and mutual funds. He/She seeks to acquire wealth and go about a simple life. Our aim was to cater to the need of this stratum of the population to help them make these decisions with a low risk-to-reward ratio. StockRC aims to bridge the information gap which is due to chunked information being spread all over the internet by different investment firms, news outlets, and finance API, and give a one-stop solution to take effective decisions without having to dig through the whole internet for these bits of information. This report describes a model that utilizes cosine similarity to perform similarity analysis on mutual funds. The model aims to recommend similar funds based on the similarity of their holdings. The mutual fund data used for this analysis was obtained from Groww, and the model operates on a user-item matrix representation of the data.

Acknowledgment

We would like to express our gratitude and appreciation towards our project guide Dr. Utkarsh Khaire whose guidance and expertise have played an important role in the completion of this project successfully. We are grateful for his unwavering commitment and dedication to providing us with the necessary resources And knowledge required for the completion of this project. We would also like to thank our academic institution for providing us with the platform and resources to undertake this project. The support and infrastructure provided by the Indian Institute of Information Technology Dharwad have been crucial in facilitating our progress and enabling us to achieve our goals. Lastly, we acknowledge all the authors and researchers whose work we have referred to. Their contributions to the field have provided us with a solid foundation for building our project.

Table Of Contents

Table Number	Title	Page Number
	Abstract	2
	Acknowledgment	3
	List of Figures	5
1	Data Collection	6
2	Forecasting Model	11
3	Application UI	15
4	Conclusion, Limitations, and Future Goals	19
	References	22

LIST OF FIGURES

Figure Number	Page Number
1	7
2	7
3	8
4	10
5	15
6	15
7	16
8	16
9	17
10	18
11	18
12	19

CHAPTER 1

DATA COLLECTION

1.1 INTRODUCTION

Data collection refers to the systematic gathering and recording of information or data for a specific purpose. It involves collecting relevant and accurate data from various sources, such as surveys, observations, interviews, experiments, or existing databases. The data collected can be in different formats, including numerical, textual, or multimedia. Data collection methods can vary depending on the nature of the research or the objectives of the organization. The collected data serves as the raw material for analysis, interpretation, and decision-making, enabling organizations to gain insights, measure performance, identify patterns, and make informed choices.

1.2 GROW DATASET

Groww[1] is an online investment platform in India that allows users to invest in mutual funds, stocks, and other financial instruments. It provides a user-friendly interface, research tools, and educational resources to help individuals make investment decisions. Groww has gained popularity in the Indian market and has a significant user base. If you are referring to Groww, you can explore their website and resources to learn more about how to grow your investments using their platform. We have collected Mutual fund information with the percentage of equity allocated for the particular stock from the overall pool of capital that the fund house has for the particular mutual fund along with the risk rating of that mutual fund along with the advanced ratios for analyzing a particular stock.

Advanced ratios			
Top 5	30%	Alpha	21.18
Top 20	64%	Beta	0.94
P/E Ratio	18.56	Sharpe	1.45
P/B Ratio	1.80	Sortino	1.74

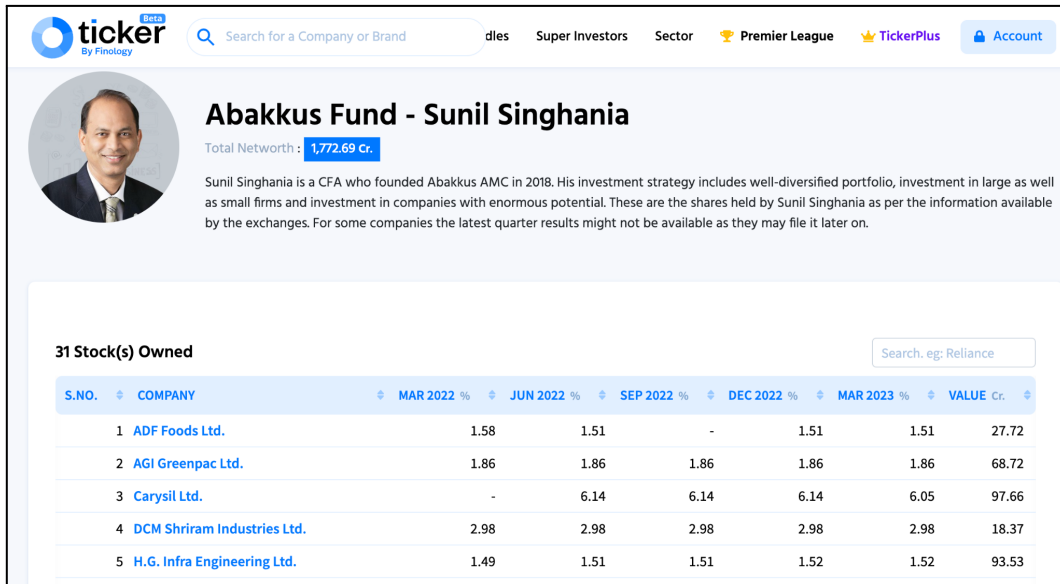
Figure 1

Name	Sector	Instrument	Assets
Reliance Industries Ltd.	Energy	Equity	8.7%
HDFC Bank Ltd.	Financial	Equity	7.9%
ITC Ltd.	Consumer Staples	Equity	5.5%
RBL Bank Ltd.	Financial	Equity	4.5%
Punjab National Bank	Financial	Equity	3.8%
IRB Infrastructure Developers Ltd.	Construction	Equity	3.8%
Bikaji Foods International Ltd.	Consumer Staples	Equity	3.6%
Jindal Stainless Ltd.	Metals & Mining	Equity	3.2%
Usha Martin Ltd.	Metals & Mining	Equity	2.9%
Just Dial Ltd.	Services	Equity	2.5%
See All			

Figure 2

1.3 TICKER DATASET

Ticker.finology[2] is a financial website that provides a range of tools and services related to the Indian stock market. It offers features such as stock screener, portfolio tracker, financial analysis, news, and educational resources. Users can search for specific stocks, access fundamental and technical data, and make informed investment decisions. icker.finology.in appears to be focused on providing financial information and analysis for the Indian market. We collected data about the amount of capital in crores allocated for the particular stock by the super investor. Data shows clear indications of super investors investing heavily in particular stocks and not much fluctuation is observed when we see the overall change in the investment strategy.



ticker Beta
By Finology

Search for a Company or Brand

dles Super Investors Sector Premier League TickerPlus Account

Abakkus Fund - Sunil Singhania
Total Network: 1,772.69 Cr.

Sunil Singhania is a CFA who founded Abakkus AMC in 2018. His investment strategy includes well-diversified portfolio, investment in large as well as small firms and investment in companies with enormous potential. These are the shares held by Sunil Singhania as per the information available by the exchanges. For some companies the latest quarter results might not be available as they may file it later on.

31 Stock(s) Owned

S.NO.	COMPANY	MAR 2022 %	JUN 2022 %	SEP 2022 %	DEC 2022 %	MAR 2023 %	VALUE Cr.
1	ADF Foods Ltd.	1.58	1.51	-	1.51	1.51	27.72
2	AGI Greenpac Ltd.	1.86	1.86	1.86	1.86	1.86	68.72
3	Carysil Ltd.	-	6.14	6.14	6.14	6.05	97.66
4	DCM Shriram Industries Ltd.	2.98	2.98	2.98	2.98	2.98	18.37
5	H.G. Infra Engineering Ltd.	1.49	1.51	1.51	1.52	1.52	93.53

Figure 3

1.4 YFinance

YFinance[3] refers to a popular Python library called "yfinance" (Yahoo Finance). It is a Python module that allows users to easily retrieve historical data, current stock prices, and other financial information from yahoo finance[4]. The finance Library provides an interface to access data such as historical stock prices, dividend data, stock splits, and more. We collected daily stock information along with past 52-week stock information with API calls

1.5 ECONOMIC TIMES

The Economic Times[4] provides comprehensive coverage of business news, economic developments, stock market updates, corporate news, and analysis. It covers a wide range of topics including finance, economy, industry, politics, technology, and more.

We had collected data about whether to buy, sell or hold a particular stock and if so then at what target price. The given data is scraped and structured into different columns so as to only get the correct information on a daily basis.

1.6 MONEY CONTROL

Moneycontrol[5] is a popular financial news and investment tracking platform in India. It provides a comprehensive range of financial information, including stock market updates, business news, investment insights, portfolio tracking tools, and in-depth analysis. We extracted daily news about the market that is to be displayed.

CHAPTER 2

FORECASTING MODELS

2.1 INTRODUCTION

Forecasting models are mathematical or statistical tools used to predict future outcomes or trends based on historical data and patterns. These models aim to provide estimates or projections of future events or values, helping businesses, organizations, and individuals make informed decisions and plans

2.2 MODEL 1

2.2.1 OVERVIEW

The popularity-based recommendation model described here utilizes a binary similarity matrix with a threshold of 0.3.

2.2.2 DATA PREPROCESSING AND MODEL IMPLEMENTATION

The input data, containing information about items and their features, is loaded into a DataFrame. The features are preprocessed, typically by converting them to lowercase or applying other necessary transformations.

Vectorization:

The TfidfVectorizer from scikit-learn[6] is used to vectorize the item features. This step converts the textual features into numerical representations that can be used for similarity calculations.

Similarity Matrix Calculation:

The cosine similarity is calculated between the feature vectors using cosine similarity from scikit-learn. The cosine similarity between x_i and x_j can be computed as:

$$\text{Similarity}(x_i, x_j) = (x_i \cdot x_j) / (\|x_i\| * \|x_j\|)$$

where $x_i \cdot x_j$ represents the dot product of x_i and x_j , and $\|x_i\|$ and $\|x_j\|$ represent the norms (magnitude) of x_i and x_j , respectively. This results in a similarity matrix that represents the similarity between each pair of items in the dataset.

Conversion to Binary Values:

The similarity matrix is converted to binary values using a threshold of 0.3. Values greater than 0.3 are set to 1, indicating similarity, while values less than or equal to 0.3 are set to 0, indicating dissimilarity.

2.2.3 Recommendation Generation:

To generate recommendations, the model identifies the item of interest for which recommendations are sought. The binary similarity values corresponding to the item of interest are retrieved. The indices of the most similar items are determined based on the binary similarity values. The names of the top similar items are retrieved from the DataFrame.

2.3 MODEL 2

2.3.1 OVERVIEW

The model leverages a collaborative filtering approach to calculate the similarity between mutual funds based on their holdings. The process involves constructing a stock correlation matrix, which measures the correlation between the assets of different funds. The higher the correlation, the more similar the investment patterns of the two funds.

2.3.2 MODEL IMPLEMENTATION

Data Preprocessing:

The mutual fund data scraped from Groww is loaded into a pandas DataFrame. The DataFrame is then transformed into a pivot table, where the rows represent the funds' names, columns represent the company names, and values represent the asset values.

Collaborative Filtering[7]:

The call filtering function takes a fund name as input. The stock correlation matrix is computed by calculating the correlation between the assets of the input fund and all other funds.

$$\text{Rating}(i, k) = \text{Avg}(i) + (\text{Sum}(\text{Sim}(i, j) * (\text{Rating}(j, k) - \text{Avg}(j))) / \text{Sum}(|\text{Sim}(i, j)|))$$

Where:

Avg(i) is the average rating given by user i. Sim(i, j) represents the similarity between user i and user j. Rating(j, k) represents the rating given by user j to item k. Avg(j) is the average rating given by user j. The sums are taken over all users j who have rated item k. The correlation values are sorted in descending order to identify the most similar funds.

2.4 MODEL 3

2.4.1 OVERVIEW

The model employs cosine similarity to measure the similarity between the holdings of different mutual funds. The cosine similarity metric calculates the cosine of the angle between two vectors, representing the funds' holdings. A higher cosine similarity score indicates a more significant similarity between the holdings of two funds.

2.4.2 MODEL IMPLEMENTATION

Data Preprocessing:

The mutual fund data obtained from Groww is loaded into a pandas DataFrame. Missing values, if any, are replaced with 0 to ensure a complete user-item matrix.

Cosine Similarity Calculation:

The cosine_simi[8] function takes a fund name as input. The user-item matrix is transposed to have funds as rows and holdings as columns. Cosine similarity is calculated between the columns of the transposed matrix, representing the holdings of different funds. Cosine Similarity between Vectors A and B:

$$\text{similarity}(A, B) = (A \cdot B) / (\|A\| * \|B\|)$$

Where:

$A \cdot B$ represents the dot product (also known as the inner product) of vectors A and B. $\|A\|$ and $\|B\|$ represent the Euclidean norm (magnitude) of vectors A and B, respectively. The similarity scores between the target fund (input fund) and all other funds are obtained.

Similarity Scores and Recommendations:

The similarity scores are sorted in descending order.

The top 10 similar funds, excluding the target fund itself, are selected based on the highest similarity scores. The selected funds and their similarity scores are returned as the output.

Results and Findings:

The model output provides the top 10 similar funds, excluding the input/target fund, based on their holdings' cosine similarity. Each recommended fund is accompanied by its name and the corresponding similarity score.

CHAPTER 3

APPLICATION UI

3.1 HOME PAGE

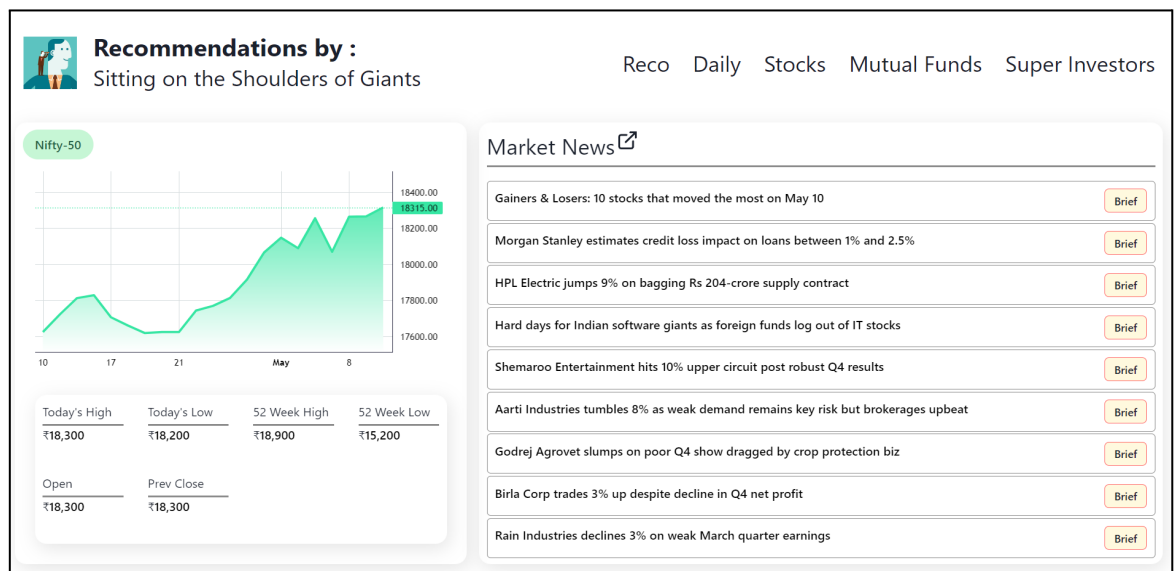


FIGURE 5

3.2 DAILY NEWS

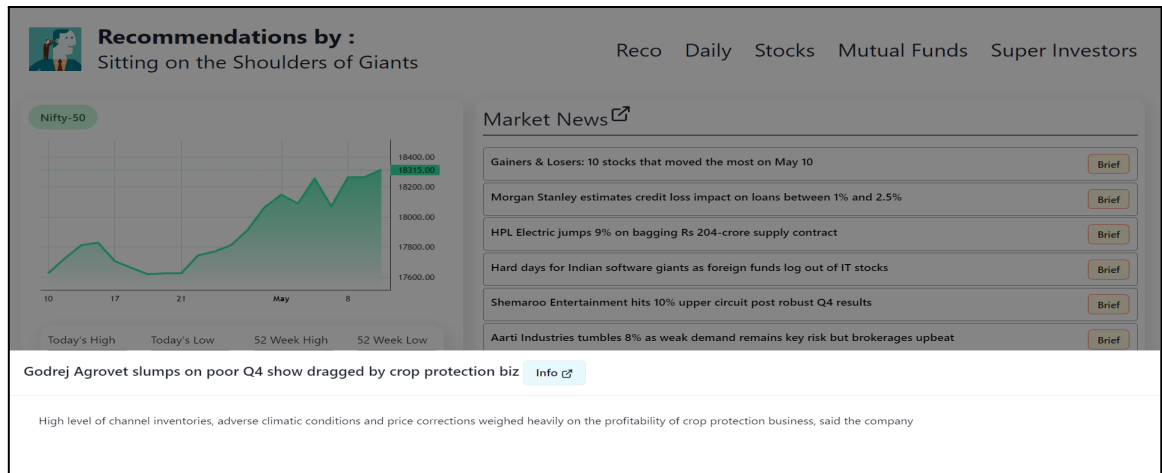


FIGURE 6

3.3 DAILY BUY/SELL/HOLD SIGNALS

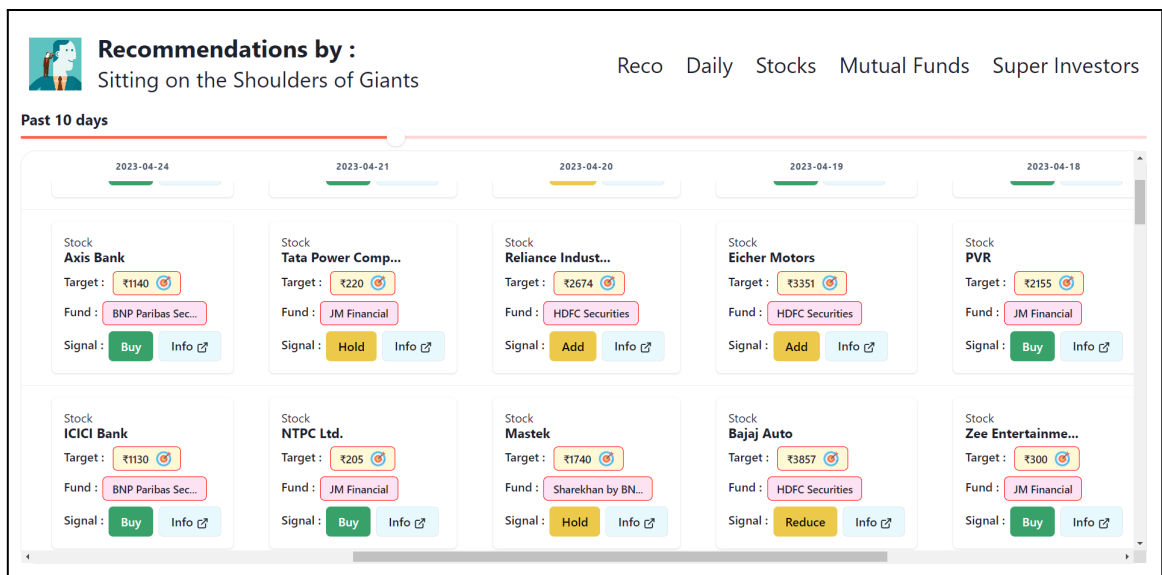


FIGURE 7\

3.4 POPULARITY-BASED STOCK RECOMMENDATION

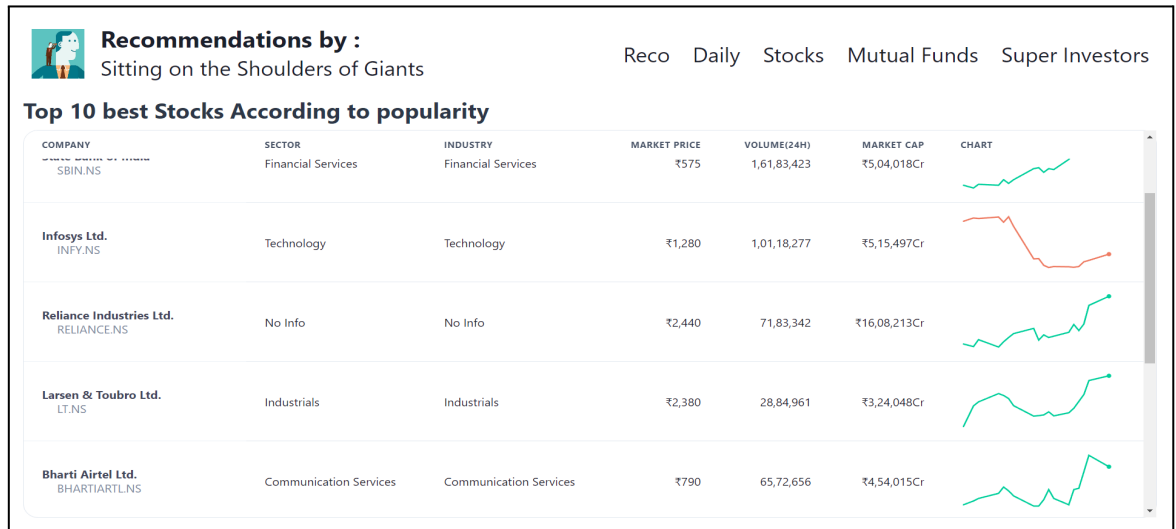


FIGURE 8

3.5 VOLUME-BASED STOCK RECOMMENDATION

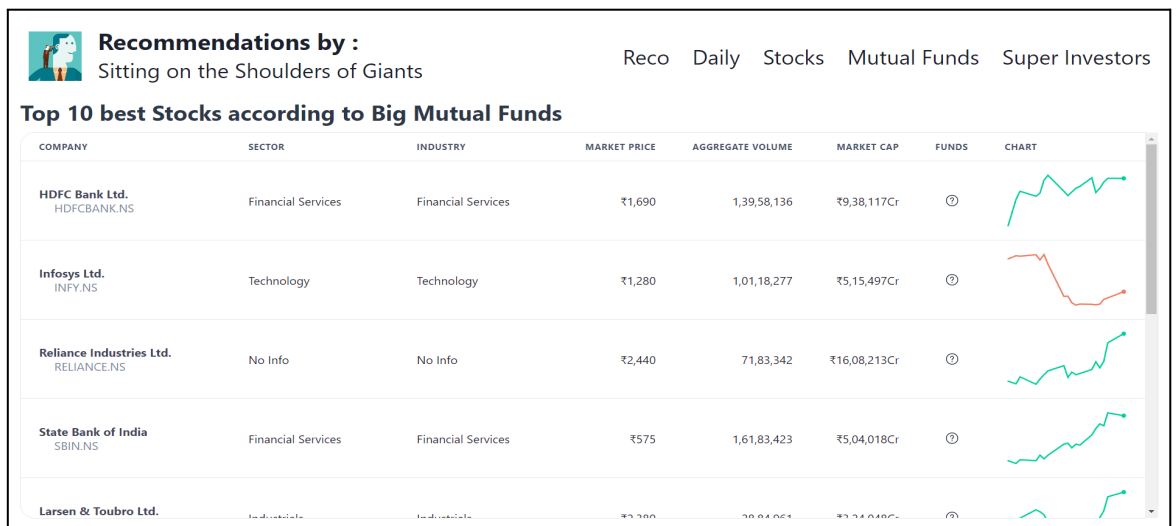


FIGURE 9

3.6 BASED ON SUPER-INVESTORS

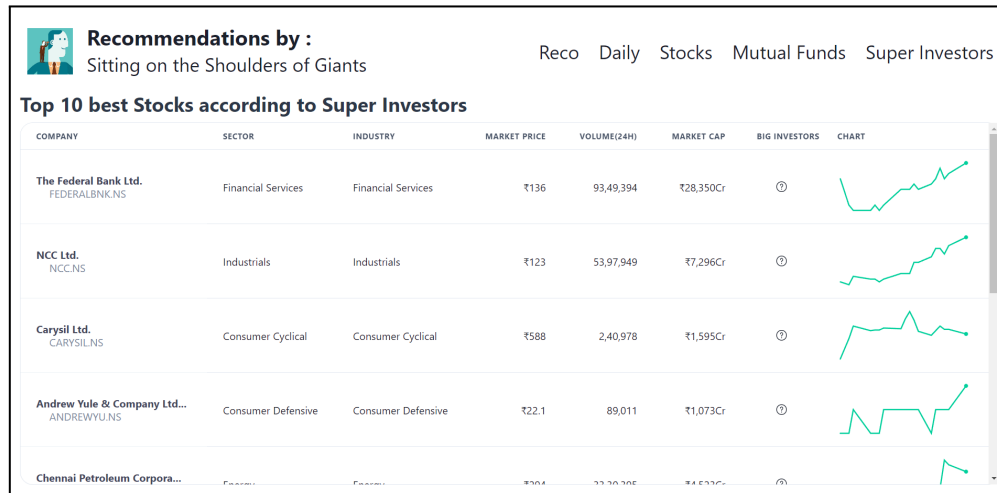


FIGURE 10

3.7 CORRELATION AND COSINE BASED RECOMMENDATION

Get me most similar stocks of the given stock

Housing Development Finance Corporation Ltd.

Model 1

Description

Correlation is another measure of similarity that is commonly used in recommendation systems. Correlation measures the linear relationship between two variables. In the context of a recommendation system, the variables represent the user's past behavior and the items being recommended. The correlation between the two variables can be positive or negative, and the closer the correlation is to 1 (in absolute value), the stronger the relationship between the variables is considered to be.

Model 2

Description

Cosine similarity is a measure of similarity between two vectors in a multi-dimensional space. In the context of a recommendation system, the vectors represent the user's past behavior (e.g., items they have viewed or purchased) and the items being recommended. The cosine similarity between two vectors is calculated as the cosine of the angle between them. The closer the cosine similarity is to 1, the more similar the two vectors are considered to be.

FIGURE 11

Results :	Results :
• Dodla Dairy Ltd	• ITC Ltd.
• Jamna Auto Industries Ltd.	• Coal India Ltd.
• Sula Vineyards Ltd.	• Hero Motocorp Ltd.
• Gokaldas Exports Ltd.	• Power Grid Corporation Of India Ltd.
• Housing & Urban Development Corporation Ltd.	• Bajaj Holdings & Investment Ltd.
• Bank of India	• Indian Energy Exchange Ltd.
• Housing Development Finance Corporation Ltd.	• HDFC Bank Ltd.
• Sheela Foam Ltd.	• Reliance Industries Ltd.
• Sagar Cements Ltd.	• Kotak Mahindra Bank Ltd.
• Fine Organic Industries Ltd.	• Central Depository Services (India) Ltd.

FIGURE 12

CHAPTER-4

CONCLUSION, LIMITATIONS & FUTURE GOALS

4.1 CONCLUSIONS

The cosine similarity-based model presented in this report offers a useful approach for identifying mutual funds with similar investment patterns. By utilizing cosine similarity to measure the similarity of holdings, the model provides recommendations that align with the investment preferences of investors. However, it is important to note that this model solely focuses on the holdings of the funds and does not consider other critical factors such as fund performance, risk, or management. Therefore, investors should conduct comprehensive research and consider multiple factors before making investment decisions.

The collaborative filtering model presented in this report offers a valuable tool for investors seeking mutual funds with similar investment patterns. By

analyzing the correlation between the assets of different funds, the model enables investors to identify funds that align with their investment preferences. However, it is important to note that this model is solely based on the holdings of the funds and does not consider other factors such as fund performance, risk, or management. Therefore, investors should conduct comprehensive research and consider multiple factors before making investment decisions.

4.2 LIMITATIONS AND FUTURE GOALS

The model's performance heavily relies on the quality and relevance of the input data. It assumes that the provided mutual fund data accurately represents the holdings of the funds. The model only considers the correlation between the assets of the funds and does not account for other important factors such as fund objectives, risk profiles, or historical performance. Future enhancements could include incorporating additional features and factors such as fund performance metrics, risk measures, and expense ratios to provide more comprehensive recommendations.

The model could be further refined by incorporating machine learning algorithms to analyze historical data and predict future trends in fund holdings. Overall, this collaborative filtering model serves as a useful starting point for investors looking to identify mutual funds with similar investment patterns. However, it is crucial for investors to conduct their due diligence and consider a wide range of factors before making investment decisions.

The model assumes that the mutual fund data accurately represents the holdings of the funds. Inaccuracies or missing data may affect the similarity analysis. The model solely relies on the cosine similarity metric and does not incorporate other factors, such as fund objectives, risk profiles, or historical performance, which are crucial in fund selection. Future enhancements could involve incorporating additional features, such as fund performance metrics and risk measures, to provide more comprehensive recommendations.

Machine learning techniques, such as dimensionality reduction or clustering algorithms, could be incorporated to enhance the model's performance and identify more nuanced patterns within the data. Overall, the cosine similarity-based model serves as a valuable tool for investors seeking mutual funds with similar investment patterns. However, investors need to consider a

broad range of factors and conduct thorough due diligence before making investment decisions.

One key limitation of using a binary similarity matrix with a threshold of 0.3 is that it simplifies the notion of similarity between items. By applying a binary threshold, we group items into two categories: similar and dissimilar. This approach may lead to some loss of information, as it does not account for the varying degrees of similarity that may exist between items.

Binary similarity matrices with a fixed threshold lack the ability to capture subtle differences in item similarities. While they can identify items that are relatively more similar to each other, they may miss out on items that have moderate levels of similarity. As a result, the recommendations provided by this model may not be as precise or tailored to individual user preferences.

REFERENCES

[1] Dass, Chiranjeev, and Mirfas Moideen. "A research study about consumer awareness and purchase intention towards mutual fund and stock market investment app 'GROWW'in India." (2020).

[2] *Market dashboard*. <https://ticker.finology.in/>. (n.d.).
<https://ticker.finology.in/market>

[3] I. Bordino, N. Kourtellis, N. Laptev and Y. Billawala, "Stock trade volume prediction with Yahoo Finance user browsing behavior," 2014 IEEE 30th International Conference on Data Engineering, Chicago, IL, USA, 2014, pp. 1168-1173, doi: 10.1109/ICDE.2014.6816733.

[4] Karmakar, Madhusudan. "Stock market volatility in the long run, 1961-2005." *Economic and Political Weekly* (2006): 1796-1802.

[5] Saumya, Sunil, Jyoti Prakash Singh, and Prabhat Kumar. "Predicting stock movements using social networks." *Social Media: The Good, the Bad, and the Ugly: 15th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2016, Swansea, UK, September 13–15, 2016, Proceedings 15*. Springer International Publishing, 2016.

[6] Kramer, O. (2016). Scikit-Learn. In: *Machine Learning for Evolution Strategies. Studies in Big Data*, vol 20. Springer, Cham. https://doi.org/10.1007/978-3-319-33383-0_5

[7] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[8] Rahutomo, F., Kitasuka, T. and Aritsugi, M., 2012, October. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST* (Vol. 4, No. 1, p. 1).