International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 04 | Apr 2023 www.irjet.net p-ISSN: 2395-0072

CLUSTERING MODELS FOR MUTUAL FUND RECOMMENDATION

Aayush Shah[1], Aayushi Joshi[2], Dhanvi Sheth[3], Miti Shah[4], Prof. Pramila M Chawan[5]

[1,2,3,4] B.Tech Student, Dept. of Information Technology, VJTI College, Mumbai, Maharashtra, India
[5]Associate Professor, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---

Abstract - *The mutual fund industry has expanded significantly, providing investors with several investment options. Mutual fund information is necessary for investors to make prudent investments. Yet, novice investors may find the financial environment to be complex owing to the abundance of information. A mutual fund recommendation system based on machine learning and data analytics overcomes this issue. We have proposed a clustering models for recommending mutual funds by analyzing theories regarding mutual fund investments and returns.*

Key Words*: *Mutual funds, Clustering models, K-means, DBSCAN, Hierarchal, Agglomerative

1. INTRODUCTION

Mutual fund investing has become an integral component of portfolio management for investors and financial institutions. Yet, choosing the best mutual funds to invest in may be difficult owing to the vast number of possibilities and the complexity of the elements that affect their performance. It is essential to accurately forecast the performance of mutual funds in order to make educated investing selections. In this paper, we have described clustering models for recommending mutual fund investments. The suggested model takes into consideration a number of implicit and explicit parameters, such as expense ratios, fund manager experience, past performance, and net asset values, in order to create investment recommendations that correspond to an investor's preferences and risk profile. The models such as K-means, hierarchical clustering, and DBSCAN group mutual funds based on their comparable traits and performance. This allows the models to offer suggestions based not just on the characteristics of individual funds, but also on the performance and behavior of funds with comparable characteristics. Using cutting-edge clustering techniques, our models provides a complete solution for investors seeking to improve their mutual fund investments.

2. PROBLEM

2.1 Problem statement

To propose clustering models for recommending mutual funds. Today, there is a lack of personalized and accurate recommendations for investors due to the vast amount of data and the complex nature of mutual funds. The existing approaches are limited and may not provide a satisfactory solution for novice investors. Hence, there is a need for an efficient and reliable recommendation system that can consider the individual preferences and risk tolerance of investors to provide tailored recommendations for mutual fund investments.

2.2 Problem elaboration

With the rise of online trading platforms, retail investors now have access to a wider variety of investment options, but the sheer number of options can be overwhelming. Additionally, many investors may lack the financial expertise to evaluate the risks and returns of different mutual funds effectively.

A mutual fund recommendation system could provide personalized investment advice based on a user's investment goals, risk tolerance, and other relevant factors. However, designing an effective system would require addressing several challenges. One of the primary challenges is building a model that can accurately predict the performance of different mutual funds based on historical data. This requires identifying relevant features that are predictive of mutual fund returns and developing algorithms that can effectively learn from this data.

Another challenge is ensuring that the system can provide personalized recommendations that reflect each user's unique investment goals and preferences. This requires developing effective methods for capturing user preferences and incorporating them into the recommendation process.

Finally, it is important to ensure that the system is transparent and easy to use for novice investors. This means designing an intuitive user interface that explains the rationale behind each recommendation and provides users with the information they need to make informed decisions.

Overall, a mutual fund recommendation system has the potential to empower novice investors and help them navigate the complex world of mutual fund investments. However, designing an effective system requires addressing several technical and user-facing challenges.

3. DATA

3.1 Data collection

We acquired our data from the Value Research Online website. It is a well-established website that provides financial information and analysis to help investors make

informed decisions about their investments. The website offers a wide range of services, including mutual fund research. Additionally, the website follows strict editorial policies to ensure the accuracy and reliability of its content.

The data provided includes critical attributes for equity, debt, and hybrid mutual fund types, which are financial vehicles that investors can use to invest in the financial markets. Equity funds invest in stocks and have higher risk and return potential, while debt funds invest in fixed income securities with a fixed rate of return and lower risk compared to equity funds. Debt funds are further classified based on the duration of bonds they invest in. Hybrid funds invest in a mix of equity and debt securities, offering a balanced mix of risk and return potential.

3.2 Data preprocessing

The data contained a lot of shortcomings that needed to be dealt with before passing it to the machine learning model. Data was scattered in separate databases with different schemas. Several records had null values. Hence, data integrations along with data cleaning steps had to be performed. Thus, to make the data more disposable, following data preprocessing steps had to be applied.

1) Data integration

For separate features, data was extracted in a separate csv file. These columns were different for 3 kinds of mutual funds, i.e., Equity, Hybrid and Debt. Hence, we created a common schema was necessary to unify these records under a common dataset.

2) Feature selection

Based on the relevance of all features, only those features were selected that may help in predicting the mutual funds.

Data cleaning

a) Dealing with null values:

- Records missing critical features:

  There are several records in the dataset where critical features such as Sharpe Ratio, Standard Deviation, and Sortino Ratio are missing. It is difficult to evaluate risk involved without these features. Hence, records without these features were discarded completely.

- Records missing a non-critical value:

  Such features were filled with the average value (mean) of the whole column.

b) Dealing with duplicates: Duplicates were deleted.

c) Handling Outliers: We used graphical methods such as box plots and whisker plots to determine the outliers.

3) Feature Extraction

a) To make the data more expressive, we converted a few categorical columns with only a few values, into one hot encoded vector. Clustering algorithms usually use numerical data and raw form of categorical data might be erroneous. Hence, in order for the clustering algorithms to work more efficiently and remove any bias, we converted columns such as fund category and fund style.

b) A few new features were added to extract valuable information from the existing columns. For example, the column called 'date' was converted to 'age_in_months' by applying appropriate mathematical functions.

To work with manager_tenure, only primary manager tenure was extracted from an array of managers.

4) Exploratory data analysis

This step involved analyzing data and comparing different gestures with each other. This resulted in a correlation matrix between all the features. Using this matrix, features which values extremely correlated to each other had had to be removed in order to remove the bias. Hence columns such as NAV_latest, NAV_previous had a correlation of 1. These columns were combined to form only 1 column called NAV_latest.

5) Scaling data

Before passing the data to the next step, the data needs to be normalized or scaled so that bigger values don't skew the clustering output. All the numerical values were scaled

By implementing these steps, we can ensure that the dataset is cleaned, filtered, and transformed into a more useful format for recommendation modeling.

Finally, after performing all these pre- processing steps, the data contained attributes denoting fund type like equity debt or hybrid, fund performance metrics like expense ratio, returns and fund manager tenure, fund style like growth, value or blend and several other numerical attributes like risk factor, net asset value, standard deviation, Sharpe ratio and standard deviation.

4. CLUSTERING MODELS

We propose four clustering models:

1) K-means: It is used to cluster and partition data into groups based on similarities by minimizing the sum of squared distances between centroids and data points.

2) Hierarchical: It is used to group data into clusters in a hierarchical manner, based on the distance between data points, without needing to specify the number of clusters beforehand.

3) Agglomerative: It is a hierarchical clustering algorithm that starts with each point as a single cluster and gradually merges them into larger clusters with more points based on their similarity, until all points belong to a single cluster.

4) DBSCAN: It is a density-based clustering algorithm that groups data points together that are closely packed and separated from other clusters, based on a user-defined minimum number of points and a maximum distance between them.

The process of analyzing and clustering data involves various techniques that can assist in identifying patterns and structures within the data. One such technique is scaling the data to normalize and standardize it to ensure that different features or variables are comparable and easier to interpret.

Scaled dataset was used to implement these four types of clustering algorithms, i.e. Agglomerative, DBSCAN, Hierarchy, and K-means. The effectiveness of the different clusters formed using these algorithms was evaluated and checked against two metrics, which were inertia and silhouette.

Inertia measures the sum of squared distances between each point and its assigned centroid in the cluster. A lower inertia value indicates that the clusters are more tightly packed and well-separated, which is a desirable outcome. Silhouette score measures how well each data point fits into its assigned cluster, by comparing the distance between the point and other points in its own cluster (cohesion) to the distance between the point and points in the nearest neighboring cluster (separation).

A high silhouette score (closer to 1) indicates well-separated clusters, while a low score (closer to -1) indicates poorly separated clusters.

By comparing the results of the clustering algorithms against these metrics, it was determined which algorithm produced the most optimal and accurate clusters.

We defined hyperparameter search dictionaries for these clustering algorithms. The parameters for each algorithm was specified with ranges of possible values. Additionally, a dictionary containing a list of features was created to use in the grid search.

5. OUTPUT

For each combination of model and hyperparameters, clustering has been performed and the results are recorded. We compare these models on the basis of the silhouette score. Fig. 1 shows the top K-means silhouette scores with maximum score of 0.256 forming 2 clusters having counts of 598 and 328. Similarly Fig. 2, Fig. 3 and Fig. 4 shows the top scores for Hierarchical, Agglomerative and DBSCAN clustering models respectively along with their cluster counts.

	Data_frame	model	inertia	silhouette	Numb_clusters	Cluster_counts	model_params
0	features	kmeans	19947.192682	0.256181	2	[598, 328]	n_init_Seed: 10 n_init_met: k-means++
18	features	kmeans	19947.192682	0.256181	2	[598, 328]	n_init_Seed: 10 n_init_met: random
1	features	kmeans	17800.876799	0.243917	3	[532, 332, 62]	n_init_Seed: 10 n_init_met: k-means++
25	features	kmeans	12270.851653	0.228805	9	[230, 221, 138, 136, 74, 62, 27, 27, 11]	n_init_Seed: 10 n_init_met: random

Fig 1: Top K-means Silhouette score

	Data_frame	model	inertia	silhouette	Numb_clusters	Cluster_counts	model_params
69	features	hierarchy	0.0	0.591561	4	[923, 1, 1, 1]	centroid
80	features	hierarchy	0.0	0.507765	4	[922, 2, 1, 1]	median
91	features	hierarchy	0.0	0.507765	4	[922, 2, 1, 1]	weighted
36	features	hierarchy	0.0	0.507765	4	[922, 2, 1, 1]	complete

Fig 2: Top Hierarchical Silhouette score

	Data_frame	model	inertia	silhouette	Numb_clusters	Cluster_counts	model_params
317	features	Agglomerative	0.0	0.209734	9	[231, 222, 156, 144, 71, 62, 28, 11, 1]	Afin:precomputed, Link:, weighted
132	features	Agglomerative	0.0	0.209734	9	[231, 222, 156, 144, 71, 62, 28, 11, 1]	Afin:euclidean, Link:, complete
169	features	Agglomerative	0.0	0.209734	9	[231, 222, 156, 144, 71, 62, 28, 11, 1]	Afin:l1, Link:, ward
170	features	Agglomerative	0.0	0.209734	9	[231, 222, 156, 144, 71, 62, 28, 11, 1]	Afin:l2, Link:, ward

Fig 2: Top Agglomerative Silhouette score

	Data_frame	model	inertia	silhouette	Numb_clusters	Cluster_counts	model_params
596	features	dbscan	0.0	0.743419	2	[925, 1]	Epslon:20.0, min_samp: 31
566	features	dbscan	0.0	0.743419	2	[925, 1]	Epslon:17.95, min_samp: 31
564	features	dbscan	0.0	0.743419	2	[925, 1]	Epslon:17.95, min_samp: 25
563	features	dbscan	0.0	0.743419	2	[925, 1]	Epslon:17.95, min_samp: 23

Fig 4: Top DBSCAN Silhouette score

	Average	Maximum
Agglomerative	0.209734	0.209734
K-means	0.246271	0.256181
Hierarchy	0.528714	0.591561
DBSCAN	0.743419	0.743419

Table 1: Silhouette scores across different algorithms

5.1 Critical clustering features

In order to determine which aspects of the clustering approach were the most important, we constructed a Random Forest Classifier model.

We used hyperparameters like Gini index and entropy to identify the key features that drive the formation of distinct clusters in a clustering algorithm.

Fig. 5 shows that the most effective feature while using agglomerative clustering is 'Equity_fund_style_Growth' followed by 'Standard_Deviation' and 'Category_Equity'. Likewise, Fig. 6, Fig. 7 and Fig. 8 show the most effective features in the Hierarchical, K-means and DBSCAN methods respectively. It is clear from the observations that 'Category' columns play a major role in almost all the clustering algorithms to divide the mutual funds into clusters.

	coef_name	agglomerative_1
16	Equity_fund_style_Growth	0.090856
19	Standard Deviation	0.087569
0	Category_Equity	0.065575
5	age_in_months	0.061665
22	NAV_latest	0.059708
1	Category_Debt	0.057653
10	Return_3m	0.057534
23	NAV_52wk_high	0.057006
2	Category_Hybrid	0.056480
13	Return_3yr	0.055746

Fig 5: Most effective parameters of Agglomerative

	coef_name	hierarchy_1
1	Category_Debt	0.180827
0	Category_Equity	0.167582
4	Risk_factor	0.082131
2	Category_Hybrid	0.081078
10	Return_3m	0.062951
19	Standard Deviation	0.062780
22	NAV_latest	0.061611
11	Return_6m	0.056768
16	Equity_fund_style_Growth	0.055060
24	NAV_52wk_low	0.052420

Fig 6: Most effective parameters of Hierarchy

	coef_name	kmeans_1
0	Category_Equity	0.266076
1	Category_Debt	0.211340
24	NAV_52wk_low	0.120285
4	Risk_factor	0.119238
15	Return_10yr	0.107392
2	Category_Hybrid	0.059352
5	age_in_months	0.046060
14	Return_5yr	0.021692
8	Return_1wk	0.008813
13	Return_3yr	0.007855

Fig 7: Most effective parameters of K-means

	coef_name	dbsc_1
10	Return_3m	0.180069
11	Return_6m	0.127944
13	Return_3yr	0.122212
9	Return_1m	0.098852
20	Sharpe Ratio	0.050415
17	Equity_fund_style_Value	0.048363
8	Return_1wk	0.043744
22	NAV_latest	0.042657
14	Return_5yr	0.033811
24	NAV_52wk_low	0.026417

Fig 8: Most effective parameters of DBSCAN

6. CONCLUSION

In this study, we successfully implemented various clustering algorithms, including k-means, DBSCAN, Hierarchical, and Agglomerative, to effectively cluster mutual funds. We evaluated the performance of these algorithms using parameters such as Silhouette score and Inertia, allowing for a comprehensive comparative analysis to identify the optimal method for clustering mutual funds. Additionally, we employed the Random Forest algorithm to determine the most influential features that contributed to the clustering results. This insightful analysis revealed the order of importance of the features in the mutual fund clustering process, providing valuable insights for future research and investment decision-making.

7. FUTURE SCOPE

Developing an efficient clustering model to analyze and categorize users into distinct clusters based on the similarities found in their data points will be the next step. The suggested methods must be further analyzed to determine which amongst them gives the best result on the given dataset. The best clustering algorithm can effectively group users together based on shared features or characteristics within a given feature space. Once users are assigned to their respective clusters, a personalized and effective recommendation can be generated based on the cluster to which the user belongs. Importantly, this recommendation is tailored while taking into careful consideration the unique constraints and limitations that apply to each user, ensuring that it aligns with their

preferences, requirements, and other relevant factors. This approach ensures that the recommendations provided are highly relevant and valuable, providing users with a superior experience while accommodating their specific needs and constraints.

Furthermore, a sophisticated recommendation system can be built that takes into account individual investor characteristics such as investment horizon, risk profile, investment type, minimum investment and so on to recommend the best possible mutual fund schemes to that particular investor that can aid novice as well as experienced investors in choosing the best scheme to invest in out of the thousands available today.

REFERENCES

[1] Aayush Shah, Aayushi Joshi, Dhanvi Sheth, Miti Shah, Prof, Pramila M Chawan, "Mutual fund recommendation system with personalized explanations", published in International Research Journal of Engineering and Technology Volume 9 Issue 11, November 2022

[2] Pei-Ying Hsu, Chiao-Ting Chen, Chin Chou & Szu-Hao Huang, "Explainable mutual fund recommendation system developed based on knowledge graph embeddings", published in Applied Intelligence Volume 52 Issue 9 on 1st July 2022

[3] Li Zhanga, Han Zhanga, SuMin Hao, "An equity fund recommendation system by combing transfer learning and the utility function of the prospect theory", published in the Journal of finance and data science on Volume 4, Issue 4, December 2018

[4] Chae-eun Par, Dong-seok Lee, Sung-hyun Nam, Soon-kak Kwon, "Implementation of Fund Recommendation System Using Machine Learning" published in Journal of multimedia information system, Sept 30, 2021

[5] Prem Sankar Ca, R. Vidyarajb, K. Satheesh Kumarb, "Trust based stock recommendation system - a social network analysis approach", published in International Conference on Information and Communication Technologies -ICICT 2014

[6] Nusrat Rouf , Majid Bashir Malik , Tasleem Arif , Sparsh Sharma , Saurabh Singh , Satyabrata Aich and Hee-Cheol Ki, "Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions " published in MDPI, Nov 8, 2021

[7] Nghia Chu, Binh Dao , Nga Pham, Huy Nguyen, Hien Tran "Predicting Performances of Mutual Funds using Deep Learning and Ensemble Techniques " published in arXiv.org School of Statistical Finance, Cornell University archive, Sept 18, 2022

[8] K. Pendaraki, Grigorios Beligiannis, A. Lappa, "Mutual fund prediction models using artificial neural networks and genetic programming"

[9] Krist Papadopoulos "Predicting Mutual Fund Redemptions with Collaborative Filtering"

[10] Yi-ChingChoua, Chiao-TingChen, Szu, HaoHuang, "Modeling behavior sequence for personalized fund recommendation with graphical deep collaborative filtering" published in Expert Systems with Applications Volume 192, April 15, 2022

[11] Giridhar Maji, Debomita Mondal, Nilanjan Dey, Narayan C. Debnath, Soumya Sen, "Stock prediction and mutual fund portfolio management using curve fitting techniques" published in Journal of Ambient Intelligence and Humanized Computing, Jan 2, 2021

BIOGRAPHIES

Aayush N Shah, B. Tech Student, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India.

Aayushi Joshi, B. Tech Student, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India.

Dhanvi Sheth, B. Tech Student, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India.

Miti Shah, B. Tech Student, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India.

Prof. Pramila M. Chawan, is working as an Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She has done her B.E.(Computer Engineering) and M.E.(Computer Engineering) from VJTI College of Engineering, Mumbai University. She has 30 years of teaching experience and has guided 85+ M. Tech. projects and 130+ B. Tech. projects. She has published 148 papers in the International Journals, 20 papers in the National/International Conferences/ Symposiums. She has worked as an Organizing Committee member for 25 International Conferences and 5 AICTE/MHRD sponsored Workshops/STTPs/FDPs. She has participated in 17 National/International Conferences. Worked as Consulting Editor on – JEECER, JETR, JETMS, Technology Today, JAM&AER Engg. Today, The Tech. World Editor – Journals of ADR Reviewer -IJEF, Inderscience. She has worked as NBA Coordinator of the Computer Engineering Department of VJTI for 5 years. She had written a proposal under TEQIP-I in June 2004 for 'Creating Central Computing Facility at VJTI'. Rs. Eight Crore were sanctioned by the World Bank under TEQIP-I on this proposal. Central Computing Facility was set up at VJTI through this fund which has played a key role in

improving the teaching learning process at VJTI. Awarded by SIESRP with Innovative & Dedicated Educationalist Award Specialization: Computer Engineering & I.T. in 2020 AD Scientific Index Ranking (World Scientist and University Ranking 2022) – 2nd Rank- Best Scientist, VJTI Computer Science domain 1138th Rank- Best Scientist, Computer Science, India.