BID-503 Machine Learning Assignment

Performing EDA on a Dataset of your choice

Roll No: BID-21012

Dataset Chosen: Mutual Funds India Dataset

Source of Dataset:

I chose this dataset from Kaggle, an online platform renowned for its extensive collection of diverse datasets, making it a valuable resource for data scientists, researchers, and learners in various fields.

Reason for choosing this dataset:

In the recent few years, financial literacy among the Indian diaspora has been increasing at an alarming rate. At this moment, even a 13 y/o also has some knowledge about finance, at least basic ones for sure, all thanks to movies, web series on similar content, and who can forget finance influencers, lol. Most Indian people don't invest directly in the stock market, the primary mode of investment among Indians is by investing a portion of their income via SIPs i.e. Systematic Investment Plans, of which the maximum portion is Mutual Fund Folios. So, studying this particular dataset would certainly help us understand what mutual funds are, the trends in mutual fund investing, what to expect in a particular timeframe, and so on.

So, before diving deep into our dataset, let's see what mutual funds are. Mutual funds are a type of investment vehicle that pools money from investors to invest in a diversified portfolio of stocks, bonds, or other assets. Mutual funds are a popular investment option in India, as they offer a number of advantages, such as liquidity, diversification, and professional management. Let's look into our data first.

The data consists of the following entities:

a. Scheme: This gives us the name for a particular Mutual Fund Folio
b. Fund Category: This gives us more insight into whether the mutual fund is a equity fund, debt fund, hybrid fund, funds of funds/FoF, etc.
c. Type of fund: This tells us whether the fund is Closed Ended Mutual Fund or Open Ended Mutual fund, wherein 'closed-end funds' have a fixed number of shares and trade on the secondary market, while 'open-end funds' issue and redeem shares on demand and trade at NAV.
d. Benchmark: The benchmarks for a particular mutual fund folio act like a parameter to be compared against the relevant market index.
   In the dataset, the mutual fund folios are compared against benchmarks: Nifty 50, Nifty 100, Nifty Midcap 150, Nifty Smallcap 250, BSE Sensex, BSE 100, BSE 200, CRISIL Composite Bond Index, Nifty Composite Debt Index, Nifty Liquid Index, Nifty Ultra Short Duration Debt Index, Nifty Short Duration Debt Index, Nifty Medium Duration Debt Index, Nifty Long Duration Debt Index

e. Net Asset Value: NAV or Net Asset Value represents the per-unit market value of the fund's assets minus its liabilities, calculated at the end of each trading day. It reflects the true worth of each unit in the fund and is used for pricing and transacting mutual fund units.
f. CAGR: CAGR stands for Compound Annual Growth Rate. It acts like a metric that is used to evaluate the mutual fund schemes on how much returns/growth this particular mutual fund has shown thereby measuring the performance of the respective folio. In this dataset, the CAGR rate of 6 months, 1 year, and 3 years is taken.
g. Minimum Investment: This shows data about the minimum value required to invest in this particular fund in a lump sum.
h. Expense Ratio: The expense ratio of the fund is the percentage of the mutual fund's assets which are used to cover the costs of operating the fund. These costs include the fund manager's salary, marketing costs, and administrative expenses.
i. SIP Minimum Investment: This shows data about the minimum value required to invest in this particular fund via Systematic Investment Plans.

While performing EDA, the necessary libraries were imported in Python, and then the dataset was loaded.

Dataset Overview:

The dataset consisted of 2555 different mutual fund folios with their respective characteristics.
It was found that the most popular category of funds are equity and debt funds.

Handling missing values:

The data was checked for missing values, wherein the SIP investments mode, which was found to be 'Rs 1000.00' was filled in the place of those missing values in the column, and the data was stored in a new CSV file. Also, the same was done with CAGR missing values i.e. filled them with means of the respective CAGRs.

CAGR Analysis:

For the CAGR data of 6 months, the calculated mean, median, and mode were 4.03, 4.56, and 4.03 respectively.
For the CAGR data of 1 year, the calculated mean, median, and mode were 8.87, 8.87, and 8.87 respectively.
For the CAGR data of 3 months, the calculated mean, median, and mode were 7.16, 7.16, and 7.16 respectively.

Expense Ratio Handling:

For the missing values in the Expense Ratio column, expense ratio values were converted to numeric values, of which the mean was taken to be replaced with those null values.

Benchmark Column Cleansing:

The 'Benchmark' column was cleaned by removing leading/trailing whitespaces, replacing 'N/A' values with missing value indicators (NaN), and then creating a new dataset that contains only rows with valid 'Benchmark' data.

NAV Analysis:

The scheme with minimum Net Asset Value was HSBC Brazil Fund i.e. NAV 6.35
The scheme with the maximum Net Asset Value was Franklin India Flexicap Fund i.e. NAV 991.87

Top Performing Schemes:

The scheme with having the maximum CAGR of 6 months, and that too with a minimum expense ratio is Tata Infrastructure Fund.
The scheme with having the maximum CAGR of 1 Year, and that too with a minimum expense ratio is Quant Small Cap Fund.
The scheme with having the maximum CAGR of 3 Years, and that too with a minimum expense ratio is Quant Small Cap Fund.

Scatterplot Analysis:

On visualizing the scatterplots created, the overall returns for 6 months are around 5-8%, for 1 year around 18-21%, and for 3 years CAGR was around 9-13%. Here, one case of a decrease in the three-year average might be the COVID-19 Pandemic which had also affected the stock market, making it fall in a downward trajectory, thereby lowering the average return percentage.

Future Prospects:

This was just the Exploratory Data Analysis part. Furthermore, this data could be utilized for more efficacious uses too like understanding the trends of the market and refining folios catering to a better risk assessment and management, thereby allowing the investors and the fund managers to make better and informed decisions. Also, we can include the fund manager’s information like his experience, fund size, and investment strategies to get a comprehensive view of a fund’s potential.

Also, using Machine Learning algorithms like regression or time series analysis, we can build a predictive model using the given parameters. If we use Natural Language Processing or NLP, we can also perform a sentimental analysis by analyzing news articles or social media sentiments related to the market. Also, we can understand the investor's behavioral pattern of investment, and risk-taking abilities, on which using machine learning approaches, we can create personalized investment recommendations for the investor.