Analysis of Mutual Funds

CMPT 733: Programming for Big Data

A study on mutual funds in India to recommend good schemes and understand factors affecting scheme performance over the short, medium and long terms

Abhinav Sood	Abhishek Arora	Deepak Nellurvalappil
sood@sfu.ca	aaa113@sfu.ca	dnellurv@sfu.ca

Table of Contents

Motivations and Background

By definition, a mutual fund is an investment vehicle that is made up of a pool of funds collected from many investors for the purpose of investing in securities such as stocks, bonds, money market instruments and similar assets. The power of mutual funds lies in that "fund managers", who invest the fund's capital and attempt to produce capital gains and income for the fund's investors, manage them. This enables small investors to invest in professionally managed, diversified portfolios of equities, bonds and other securities, which would otherwise be very hard to create with limited knowledge and a small amount of capital.

There are various factors to consider while making an investment decision:

- *Desired Income*: A regular current income or long-term capital gains or tax benefits, etc.
- *Risk Appetite*: A low-risk, low-gain conservative portfolio or high-risk, high-gain volatile portfolio
- *Time Horizon*: Liquidity concerns in the short, medium and long terms
- *Fund type*: Capital appreciation in equity fund or mixed investment in stocks/bonds using balanced fund?
- *Fund category*: diversified or narrow? Blue chips or energy?
- *Size of the fund*: assets managed by the fund
- *Historical returns*
- *Benchmarks and benchmark performance*

When we consider the aforementioned points, it becomes easy to see why still only less than 10% of Indian households invest in mutual fund schemes, despite them being fairly well regulated by Association of Mutual Funds in India and being managed by professional fund managers. According to a 2013 Boston Analytics research report, this low number in a big market like India results from perceived high risk in investments and lack of information on how mutual funds work. This motivates us for undertaking this project.

Problem Statement

With our project, we aim to perform an analysis of mutual fund schemes in India to recommend good investment options. We would also educate our users about the factors and the degree to which these factors affect the mutual fund performance in different time horizons ranging from 1 month to 5 years.

Even with such a low investor engagement as seen above, the average assets under management across all the asset management companies amounted roughly to the tune of USD 184.6 billions [1][2][3] and amounts in sales mobilized by all schemes roughly to the tune of USD 390 billions [1][2]. Therefore, pursuing such an analysis would be huge avenues for business development, growing and sustaining customers by offering the unique value add in the form of educating them, and a major incentive to the business itself in terms of developing in-house understanding of mutual fund performance dynamics.

Data Processing Pipeline

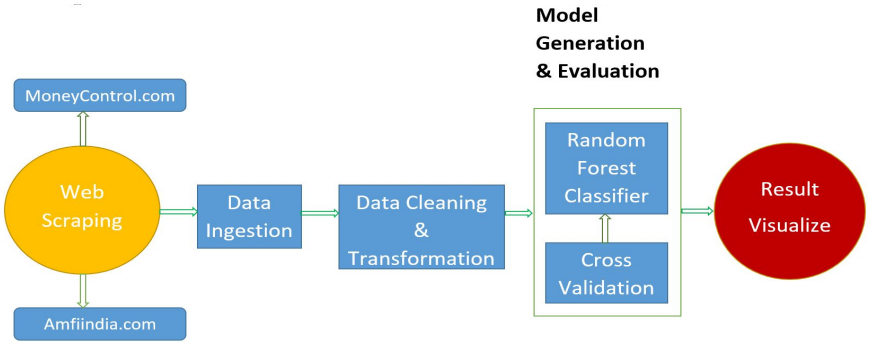
Sources of Data

We collect the data about mutual fund families and individual schemes operating in India from two disparate data sources:

1. Association of Mutual Funds in India (or AMFI) [4] - the association of SEBI (Securities and Exchange Board of India) registered mutual funds in India of all the registered Asset Management Companies. [5]
2. MoneyControl.com [6] - largest online financial platform in India [7]

Pipeline

As the data from either of the sources is not available as a structured dataset available for download, we wrote web-scrapers to scrape latest net asset value (NAV) data from AMFI [5] which is updated daily and detailed financial information about the fund as well as returns history from MoneyControl.com



The source data is collected and stored as CSVs. The fund family and fund schemes data for analysis lies in a list of fund schemes, each of which is represented as a dictionary (key-value store) of various attributes of the fund scheme. The data has to be cleaned and transformed extensively to be of any practical use. Some of the data cleaning, preprocessing and transformation steps that were performed are listed:

1. While scraping AMFI data, patterns have been identified in data to parse the structure. By going over the contents of the file [5], we observe that
   a. The first line represents the titles of columns in a ;-delimited file
   b. There are blank lines that have to be ignored
   c. There are lines with only text, and no ;-delimited values, which may represent either mutual fund scheme's type or the name of a fund family:
      i. If a single line of text is encountered before a line containing ;-delimited values it is to be interpreted as the fund family name for all funds until next such line is encountered
      ii. If two lines of text are encountered before a line containing ;-delimited values, then the first line is to be interpreted as scheme type and the second line as the fund family name for all funds until next such line is encountered
   d. Create extra fields for better representation of data to join with other dataset - namely, scheme classification, type, category, fund family, ID and a short name. All of these are derived from composite data appearing in a single field.
2. While scraping MoneyControl data, we:
   a. Collect everything as Unicode strings, and fill missing values ('NA', 'N.A.', '--', '-', None) with unicode text 'None' instead of the None keyword for the sake of consistency and ease of processing later
   b. Implemented a method `encode_risk()` to encode risk into a numeric value on a scale from 1 to 5. The higher the risk, the lower the score.
   c. Implemented a method `to_numeric()` to convert all categorical attributes as well as numerical attributes formatted as text into numerals - wrote a regular expression that handles currency, CRISIL rankings, numbers formatted with commas, etc. and works well with decimals and signs.

3. AMFI dataset gives us 12935 individual fund schemes. We restrict our analysis to the top 10 fund families with the largest assets under management (the individual schemes in these families may still have very little assets under management) as listed on MoneyControl.com AMC Asset Monitor. Therefore we have only 1296 schemes from 10 fund families in the final dataset.
4. We define and compute additional normalized metrics for each such as:
    a. Risk score based on MoneyControl risk rating - between 0 and 1
    b. CRISIL Rating (accreditation agency rating) depicting trustworthiness of fund scheme - between 0 and 1
    c. Ratio of AUM for fund scheme relative to AUM of fund family to which it belongs depicting the confidence that the fund family has in the scheme - between 0 and 1
    d. Fund performance relative to category performance - either 0 (if fund performance less than category performance) or 1 (if fund performance greater than category performance) (calculated for each time horizon)
    e. Volatility in fund scheme's category as ratio of category's worst to best performance - between 0 and 1 (calculated for each time horizon)
5. We use an Imputer for preprocessing and replace any NaN or missing values with 0.

Label Generation

We compute Expectation for each time horizon (1m, 3m, 6, 1y, 2y, 3y, 5y) based on the 5 metrics stated above to get an expected value between 0 and 1. We round the value to either 1 representing a good investment option or 0 otherwise.

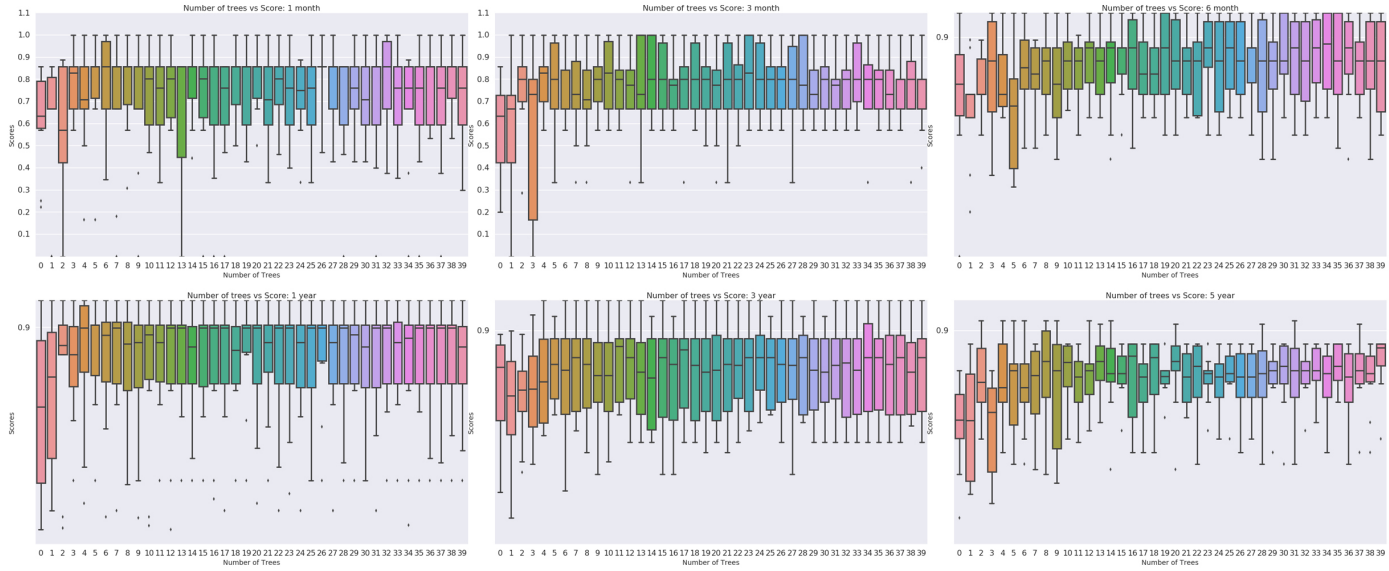
Feature Selection and Classification

We use the following as the features for our binary classification task:

1. Scheme Risk
2. CRISIL Rating
3. Fund Family AUM
4. Scheme AUM
5. Latest Net Asset Value
6. Minimum Investment for scheme
7. Latest Dividend
8. Scheme Bonus
9. Fund Return in any time frame
10. Category Average Return in any time frame

On further analysis, we realize that only a few funds have paid out any dividends or issued bonuses, therefore these are unimportant features and can be eliminated.

We separate ~23% data as test data and ~77% as training data. We perform 10-fold cross validation on Random Forest Classification with 1 to 40 trees in the forest, on training data for each individual time frame. We use box plots to visualize the results of cross-validation and pick an ideal estimator size (for example, index 26 has the most precise score for 1 month plot, index 24 is ideal in the 5 year box plot).



We train our Random Forest Classifier with the ideal estimator on the training data and check it's performance on test data by predicting labels and comparing them with the pre-assigned labels. We also generate feature importance charts from random forest classification to educate our users about the features to look at for any time horizon.

Results

We predicted the following funds for the given time frames, and analyze the performance of a fund (Birla SL Pure Value Fund (G)) predicted for both 3-year and 5-year terms:

Good Performing Funds for 1 year

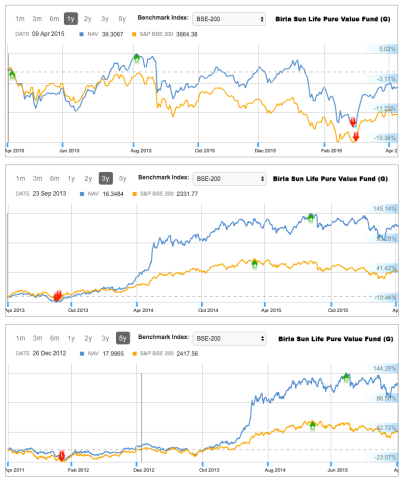
Mutual Fund Scheme	6m	1y	3y	5y
SBI Savings Fund - Direct (G)	4.2	8.8	9.4	-
Franklin (I) Savings Plus - IP (G)	3.9	8.7	9.0	9.0
Franklin (I) Savings Plus - DP (G)	4.2	8.6	9.3	-
Reliance Liquid - Cash -Direct (G)	4.1	8.4	8.9	9.0
Reliance Medium Term Fund (G)	3.9	8.4	8.8	-

Good Performing Funds for 3 years

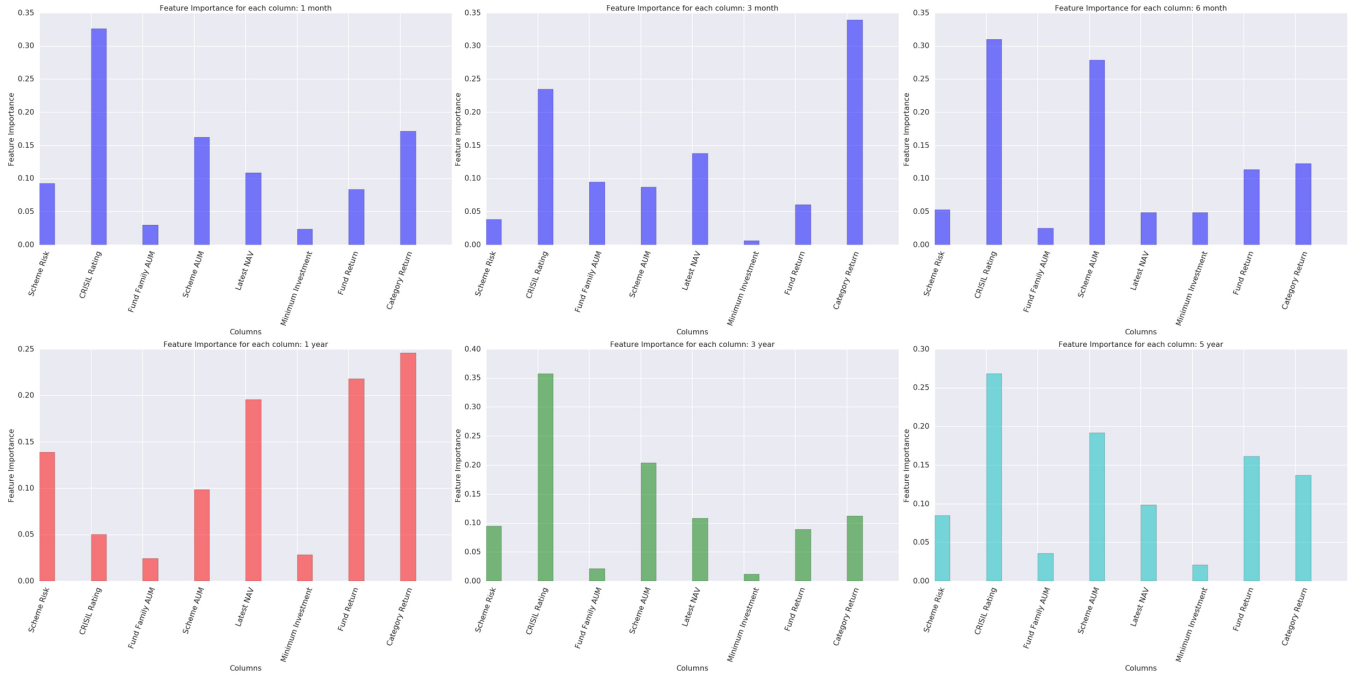
Mutual Fund Scheme	6m	1y	3y	5y
Birla SL Pure Value Fund (G)	-1.8	-2.0	30.0	17.6
Reliance Mid & Small Cap Fund (G)	-8.7	-9.7	26.0	14.5
DSP-BR Small & Mid Cap -RP (G)	-2.1	-4.1	25.6	14.7
Birla Sun Life Midcap Fund (G)	-6.4	-4.9	23.8	13.9
IDFC Premier Equity - Regular (G)	-5.2	-9.0	22.5	16.3

Good Performing Funds for 5 years

Mutual Fund Scheme	6m	1y	3y	5y
Birla SL Pure Value Fund (G)	-1.8	-2.0	30.0	17.6
SBI Emerging Busi (G)	0.1	-2.5	18.6	17.0
Reliance Mid & Small Cap Fund (G)	-8.7	-9.7	26.0	14.5
Birla Sun Life Midcap Fund (G)	-6.4	-4.9	23.8	13.9
ICICI Pru Balanced Adv (G)	-2.1	0.2	15.2	12.8



We also obtained relative feature importance for each time frame and gained insights such as the importance of CRISIL Rating (trustworthiness by an accreditation agency), Assets Under Management (AUM) and Scheme Risk increases from short term to long term, whereas minimum investment, fund and category returns usually remain similarly important:



We obtain the following classification scores:

Time Frame	Score %
1 month	96.62
3 month	96.28
6 month	94.25
1 year	98.98
3 year	94.25
5 year	95.60

Analysis and Lessons Learnt

The scores appear to be very high because of the class imbalance problem. Out of a total of 1296 data points, only these many are labeled as good, the rest are bad:

Time Frame	Good Samples	Total Samples
1 month	54	1296
3 month	52	1296
6 month	61	1296
1 year	72	1296
3 year	139	1296
5 year	107	1296

This means that if the classifier blindly assigned zeros to every data point, it would still produce a good score just because it correctly labeled bad data points as bad by chance. We can observe this by calculating the precision, recall and plotting Receiver Operating Characteristics (ROC curves). The problem can be solved in two ways - either by reducing the number of bad samples (not recommended for this particular scenario) or by increasing the good samples (which can be done by duplicating the good samples). The model trained after making these changes would perform better on unseen samples. We would like to solve the Class Imbalance problem and train a better model as a future improvement for this product.

Data Product

Our data product is available as a ready-to-execute notebook hosted on the cloud, with inline visualizations enabled so that the mutual fund scheme analytics can be visualized on the web. The webpage can be reached

at https://ec2-52-34-246-232.us-west-2.compute.amazonaws.com:8888/notebooks/MoneyControl.ipynb (password 733) where instructions can be run step by step and results can be visualized on demand.

Tools and Technology

1. Python 2.7 (programming language) with the following modules: Beautiful Soup, LXML, Requests, Numpy, Scikit Learn, MatPlotLib, Seaborn
2. Amazon Web Services (EC2, S3, I&AM)
3. Jupyter for iPython Notebooks hosted on AWS

Improvements and Future Plans

To improve our project and pursue our analysis further in the future, we would like to develop and examine models based on individual volatility measures such as the Sharpe Ration, Sortino Ration, Standard Deviation; modern portfolio theory statistics such as R-Squared, Beta, Alpha, Treynor Ratio; and upside and downside capture ratios. We have identified `MorningStar.in` as a potential source of this information (Listed under Risk and Rating for individual fund schemes, see [8] as an example). Data wrangling would be required to map unique identifiers of mutual funds between AMFI and MorningStar.in to join the datasets. We would like to solve the Class Imbalance problem that we encounter in undertaking this project.

The label generation would follow the logic for risk measurement metrics presented at Investopedia [9]. Another extension of the project could be comparing the risk assessment done as mentioned before with the risk ratings given by financial information portals like MoneyControl.com

Previous Related Work

1. G. V. Satya Sekhar. The Indian Mutual Fund Industry [internet]. Basingstoke: Palgrave Macmillan; August 2014. [Cited 2016 April 9]. Available from: http://www.palgraveconnect.com.proxy.lib.sfu.ca/pc/doifinder/10.1057/9781137407993.0001
2. Sapar, Narayan Rao and Madava, Ravindran, Performance Evaluation of Indian Mutual Funds. Available at SSRN: http://ssrn.com/abstract=433100 or http://dx.doi.org/10.2139/ssrn.433100
3. Mahajan, Nijhara, Tarani, Grover, Kumar, University of Delhi, Performance and Evaluation of Mutual Funds in India http://www.academia.edu/8085828/Performance_Evaluation_of_Mutual_Funds_in_india examines return from 15 mutual funds to find out relationship between market returns and scheme returns, and identify total and systematic risk
4. Meenakshi Garg, Dr. S.L. Gupta, A study of performance evaluation of selected mutual funds in India - available at http://shodhganga.inflibnet.ac.in/bitstream/10603/17475/13/13_summary.pdf explore the differences in performance indication by different Modern Portfolio Theory statistics, and explore correlation between index based returns and scheme returns, etc. However, the research and results are not reproducible as no code or visual references or analytics have been provided, and hence we explore some hypothesis in our project.
5. Bhatt P, Bandopadhyay A. Performance Evaluation of Schemes of Indian and International Mutual Funds: An Empirical Study of Selected Equity Large Cap Funds. Journal Of Finance, Accounting & Management [serial on the Internet]. (2011, July), [cited April 9, 2016]; 2(2): 58-70. Available from: Business Source Complete.

References

1. http://portal.amfiindia.com/spages/aqu-vol15-issueI.pdf
2. http://www.amfiindia.com/research-information/amfi-newsletter
3. http://www.amfiindia.com/research-information/aum-data
4. http://www.amfiindia.com/
5. http://portal.amfiindia.com/spages/NAV0.txt
6. http://www.moneycontrol.com/mutualfundindia/
7. http://www.moneycontrol.com/cdata/aboutus.php
8. http://www.morningstar.in/mutualfunds/f0gbr06ram/sundaram-growth/risk-ratings.aspx
9. http://www.investopedia.com/articles/mutualfund/112002.asp