Rohan Khameshra
rohankhameshra96@gmail.com

Mutual Fund Analysis

Ranking and effect of macroeconomic factors on the performance

OVERVIEW

The report outlines the analysis on the data for the eight mutual fund houses in India from the period of January 2013 onwards. The purpose of the analysis is to arrive at a concrete methodology to evaluate the performance of the eight mutual fund houses based on several mutual fund attributes and some macroeconomic factors. The ranks based on the Net Asset Value (NAV) and Asset under Management (AUM), provide a criterion to determine the performance of a mutual fund company compared to its competitors.

INTRODUCTION

The model developed for evaluating the performance along with a high-level analysis of inherent relationship of NAV and AUM with the several mutual fund attributes is utilized to design an optimal visual representation of the relationship amongst the variables.The project starts with the data scraping process and then pre-process the data set to arrange, clean and equalise the no. of data points for each of the variables. Since different variables are in different format ,to perform analysis they were converted to same format. Feature selection was done using random forest and linear regression. Then the model was trained and fine tuned , after which the feedback was checked and then trained again until the best possible prediction could be generated.

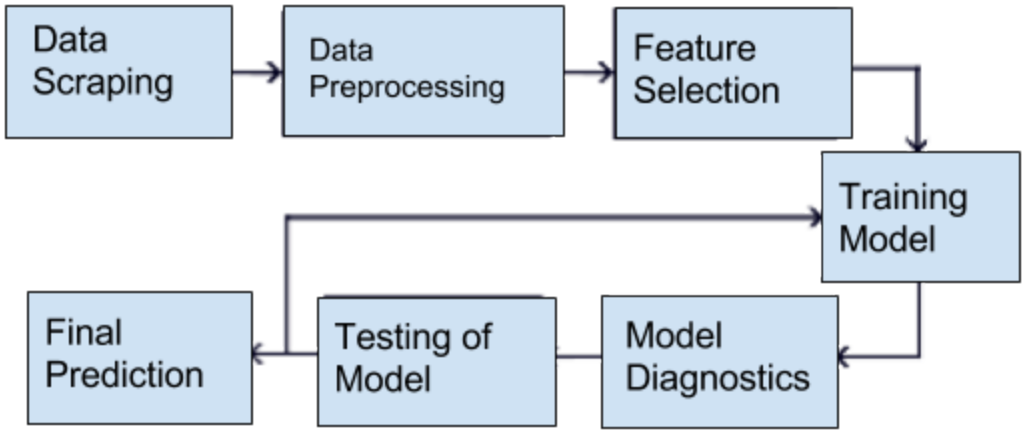


Fig 2 : Flow Chart for the Entire Methodology

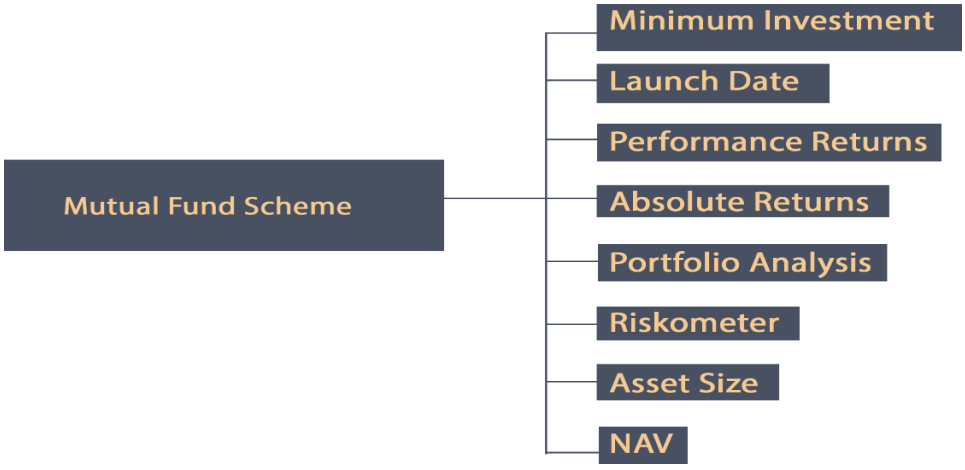
DATA SCRAPING

Websites Scraped: “http://www.moneycontrol.com/”, AMFIIndia - “https://www.amfiindia.com/”, NSE India - “https://www.nseindia.com/”, World Bank Data - “http://www.worldbank.org/”, RBI - “https://www.rbi.org.in/”, MOSPI - “http://www.mospi.gov.in/”, Open Government Data - “https://data.gov.in/”

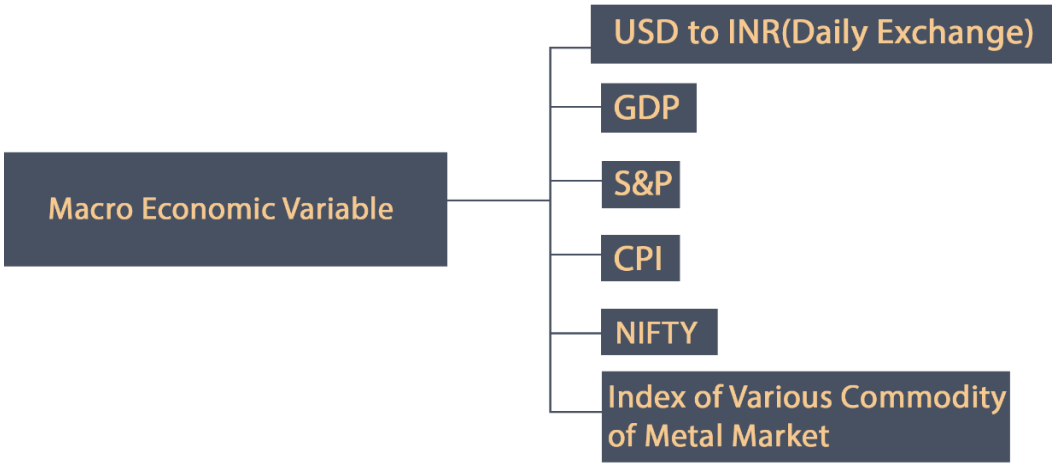
The website “Money Control” has been crawled corresponding to each Fund Group. By using Fund Group names, we extracted different attributes of a particular Fund Policy like - “AUM”, “Performance”, “Portfolio Analysis”, etc. The scrapping has been done using the “Beautifulsoup” tool in the Python framework. In order to get the AUM mechanize library has been used.

Scrapping resulted in 145 features and 1231 rows corresponding to Fund Policies. The data thus obtained contains features including 195 etc. Data present in line charts were also scrapped providing the features like NAV and Index of various Commodity or Metal Markets.

DATA DESCRIPTION

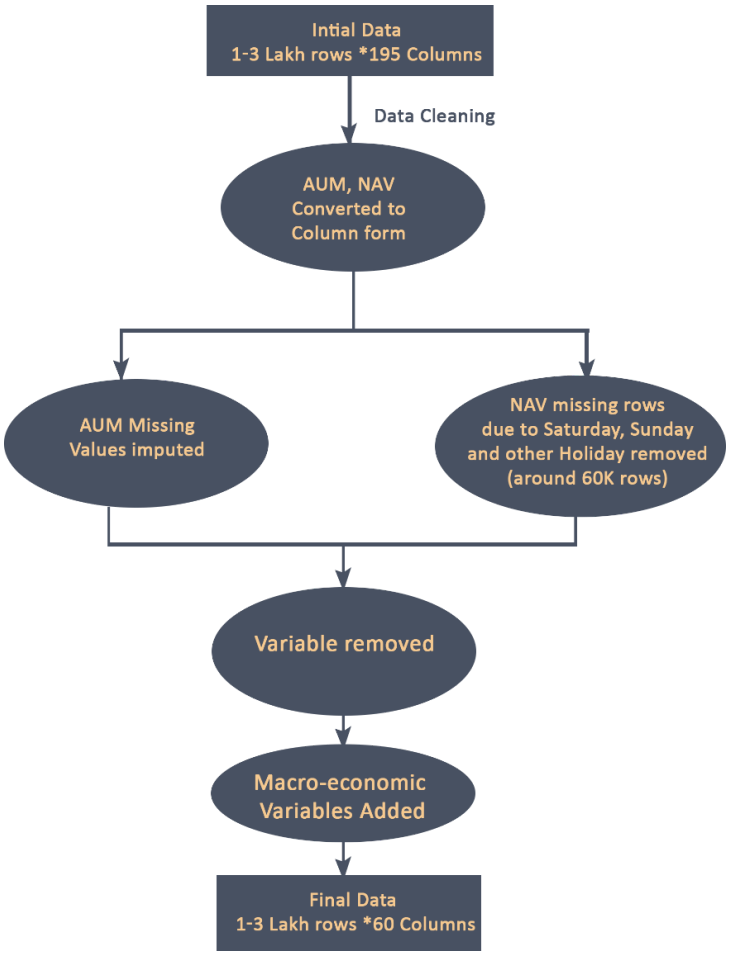


List of variables after Data scraping



List of macroeconomic Variables

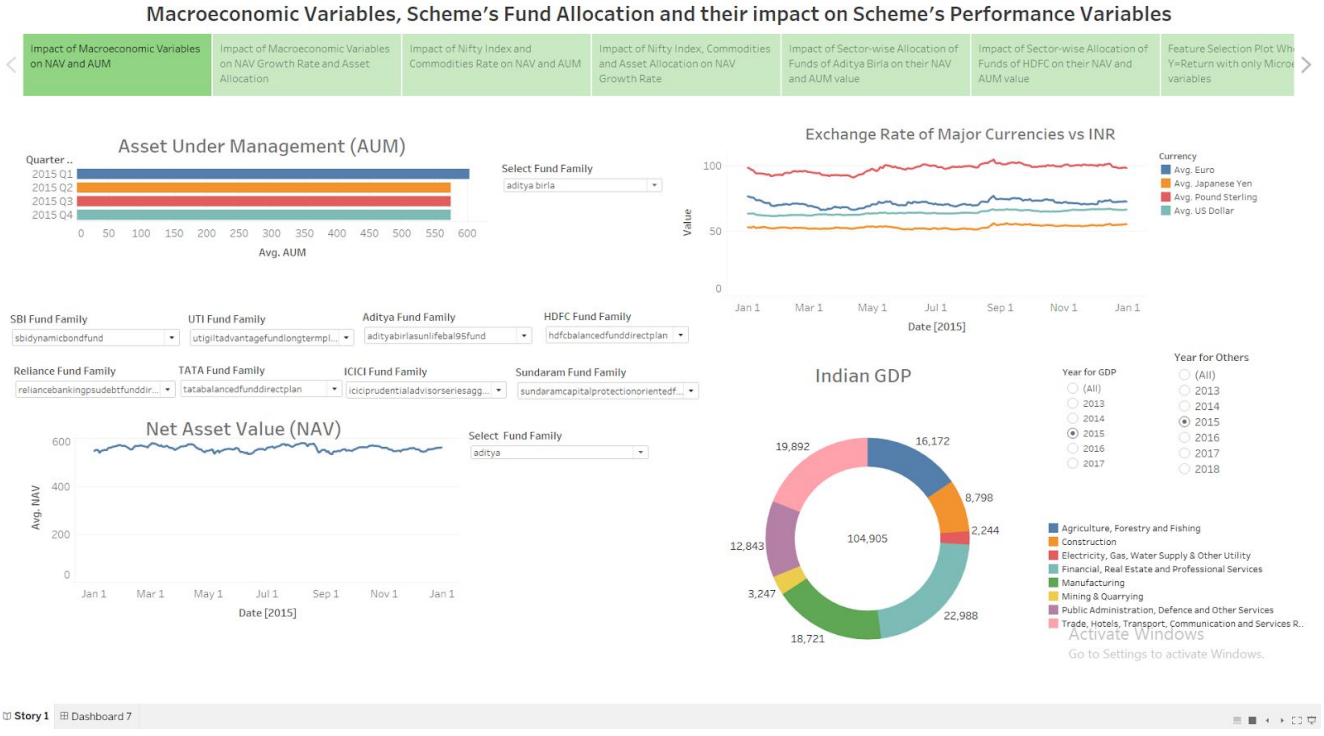
DATA CLEANING AND PREPROCESSING



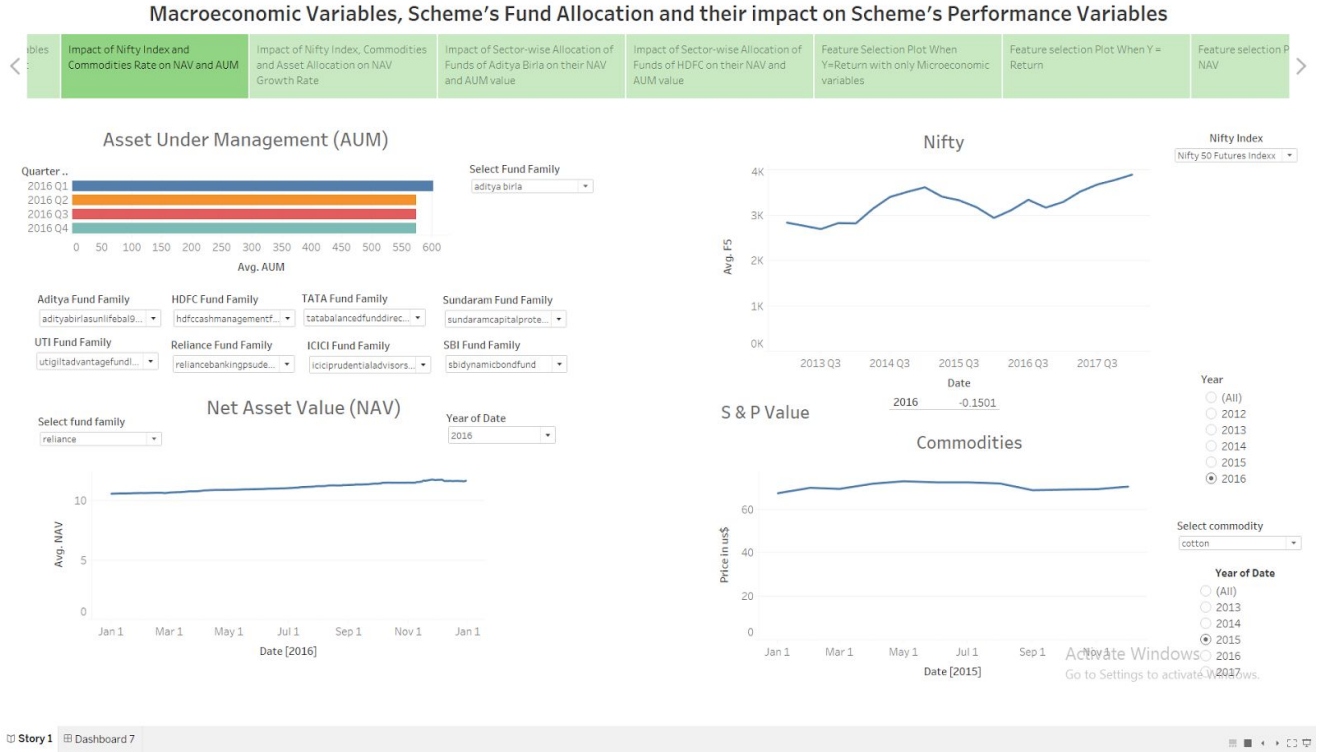
In order to find the effect of the volatility of the market on the performance of the mutual fund schemes, VIX NSE data was extracted from the National Stock Exchange site. The initial VIX contains many variables such as the closing level, the opening levels, the highs and the lows. The VIX High, low, open, previous close were removed. VIX returns were calculated using VIX close and all the remaining VIX variables were removed. Exchange rate of the Euro, Yen, Sterling were removed because on analysing the mutual fund schemes in detail it was found that none of the exchange rates affect the Indian market significantly. Moreover, due to the arbitrage free condition to hold true, all the exchange rates are related to each other. So only one Exchange rate i.e US-INR exchange rate was kept in the data while the rest of the exchange rates were removed. The initial data also contained different Commodity Prices like Aluminium, copper, cotton, crude oil, gas, gold, iron, lead, nickel, zinc. In order to mimic the effect of all the commodities variables, instead of using the individual commodities a commodity index is used. The commodity index used is MSE commodity index. % Change in NAV High and Low over a 52 week period was calculated. Last dividend column was changed to numeric. More macro variables were added namely: "Short.Term.Rate", "IIP.Growth.Rate", "Net Capital Inflows" and "Inflation.Rate". The detail of the data cleaning process is illustrated in the following infographics.

DATA VISUALIZATION

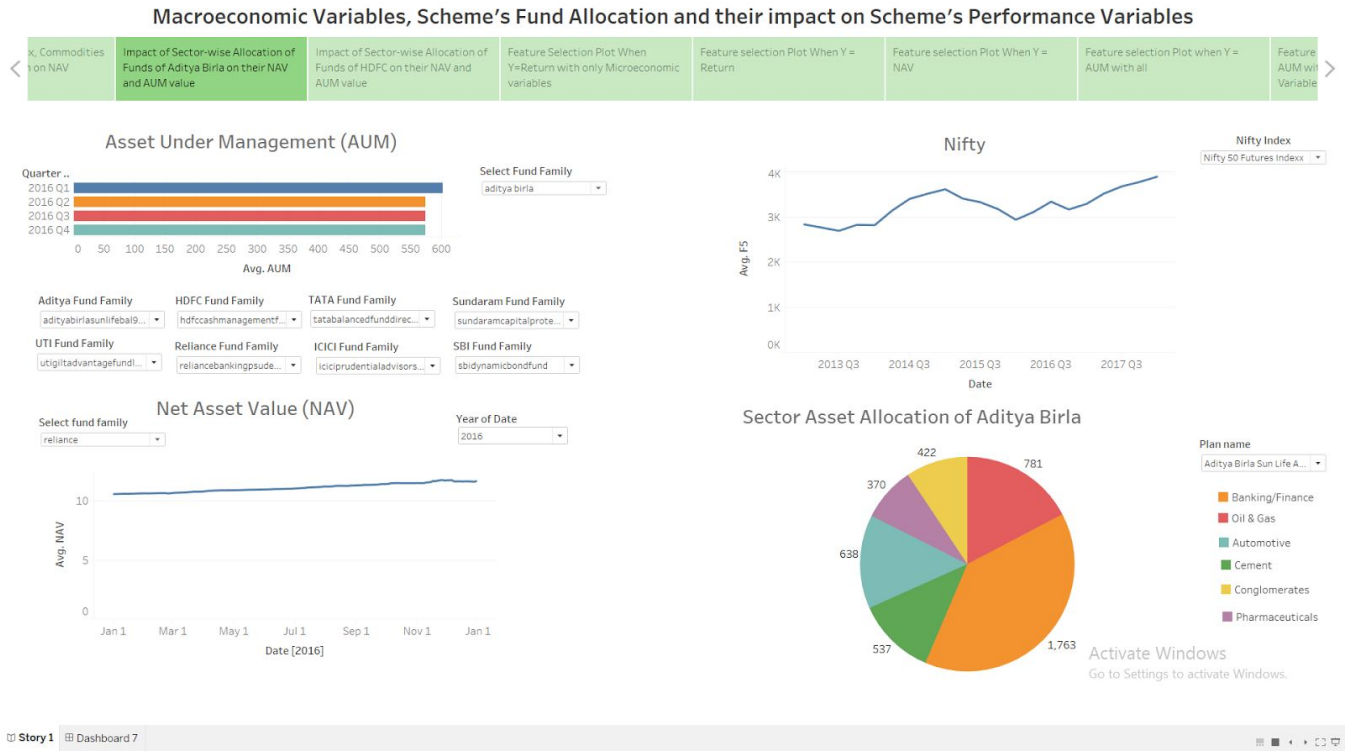
For the purpose of the data visualization,the project employes Tableau. Tableau is a advance desktop application which is used for making interactive data visualization.



The above tab shows four figures that are of AUM, NAV, GDP of India(in -US$) and Exchange Rate (US$, Euro, Yen, Pound to INR). There are two drop down menus on the right hand side to select the year, one for the year and one for the other three graphs. Moreover, for AUM and NAV figures, we can also select the mutual fund house and the required mutual fund scheme for which the graph is to be seen. The GDP is shown in a pie chart in which GDP of individual sectors is shown and the total GDP is shown in the centre of the pie chart. NAV and the exchange rate has been given on a continuous line chart against date. The purpose of this tab is to see how macroeconomic variables like Exchange Rate and GDP of the country affects the performance variables like AUM and NAV of a particular scheme.



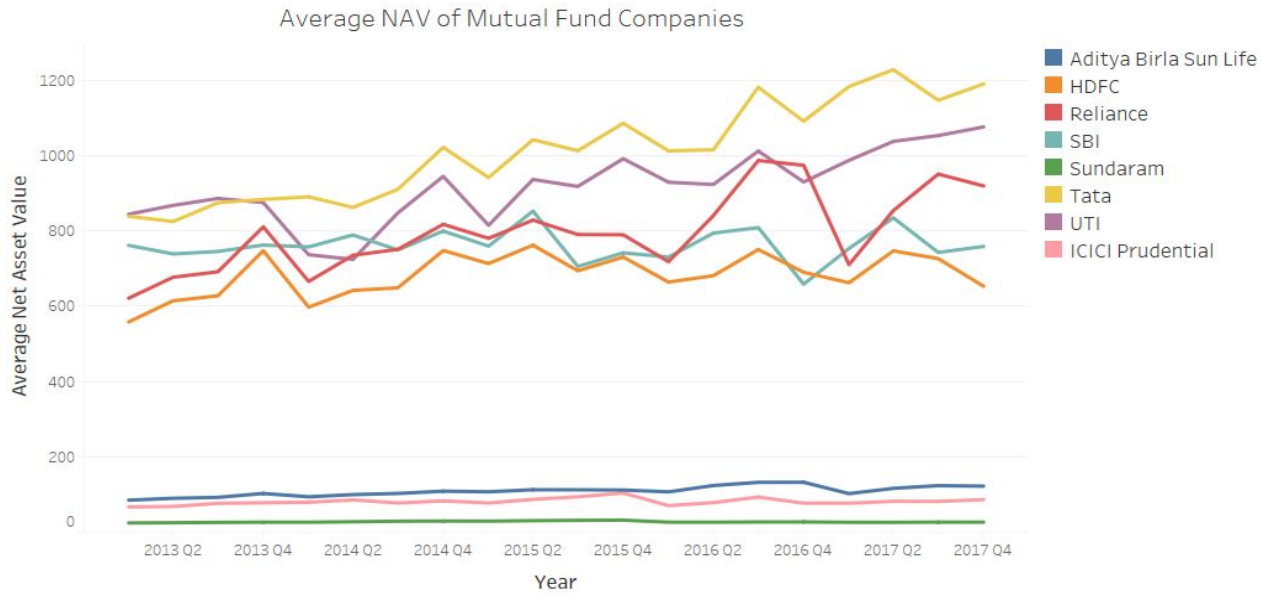
The above tab represents five figures that are of S&P Index, Nifty Price, Different Commodity Index, NAV and AUM values. In this tab also there is an option to choose year from the drop down menu as desired. In Nifty index and Commodities graph, there is one more drop down menu to select its particular type. And for AUM and NAV, we can also select the fund house and scheme type as required. This tab gives us an insight as to how different commodity prices and Indexes of the two major stock markets of India and America affect the AUM and NAV variables of each scheme.



The above tab consists of five figures that are of S&P Index, Nifty, Industry wise Fund Allocation of particular scheme, NAV Growth Rate and Asset Allocation. There are different menus to choose years for different plots. AUM data is given on a quarterly basis, while NAV is month wise.In left middle part there is option to choose between different Fund Houses and then a particular scheme. This tab mainly gives us how sector wise allocation of fund under a particular scheme affects major performance variables like NAV, AUM.

EXPLORATORY ANALYSIS

With a focus on summarizing and visualizing the important characteristics of a data set, exploratory data analysis assists in understanding the data's underlying structure and variables, developing intuition about the data set and deciding how it can be investigated with more formal statistical methods. After a detailed exploratory analysis, the project gathered some results



One interesting thing to note while considering average NAV across all mutual fund schemes for a particular mutual fund house, we find that around Q3 2016, there was a large peak for all mutual fund companies except Aditya Birla Sun Life, Sundaram and ICICI which had a comparatively small peak as these fund houses have asset size lower in compared to the other fund houses. Also, around Q1 2017, average NAVs of these same companies went crashing down as the overall market went for a dive at the same time. Also, smaller peaks were found around Q4 2014 and Q4 2013. Let us look at some sector wise pie charts for one of the 8 mutual fund houses.

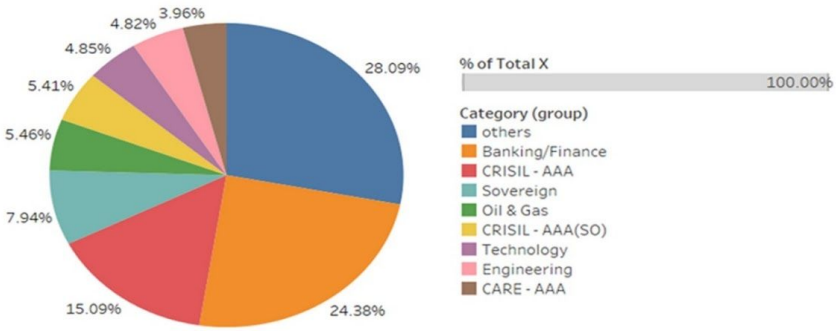
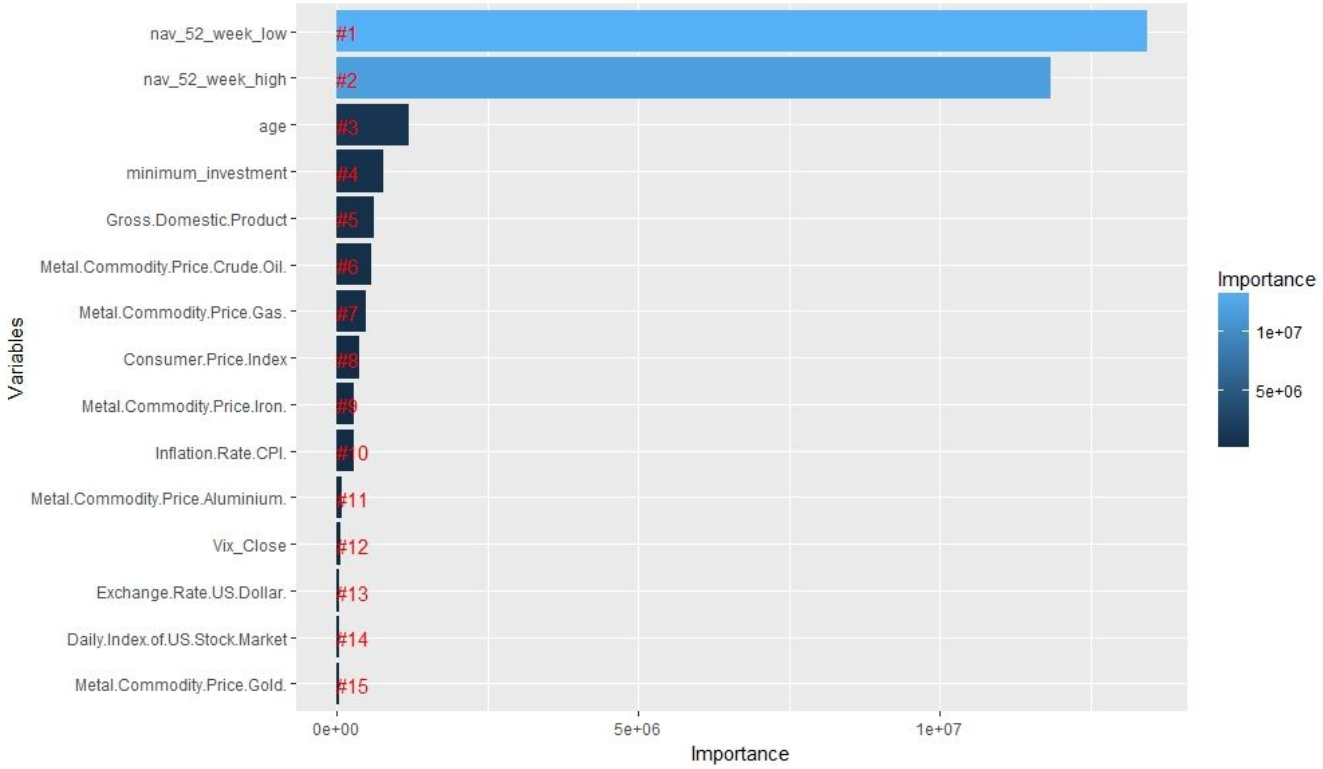


Fig 8 : Sector wise Investment for HDFC funds

FEATURE SELECTION

Random Forest Variable Selection:

Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, design to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. For regression trees it is variance. Thus when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree. For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure.



In data considered return is calculated for every Mutual Fund Scheme for all the eight fund families and 25 features are selected from the scraped data and random forest model is applied on 3 different dependent variables like Return, NAV, AUM. Later the 3 random forest feature selection method mentioned above is applied on 3 different dependent variables.

Feature selection using Linear Regression

The data consisted of mutual fund schemes belonging to different fund houses. All the fund houses followed the same financial norms but their management styles vary a lot. The different investment strategies determine the differing returns on their mutual funds. Due to this, different schemes gets impacted by different parameters. Therefore, in order to predict the performance of the schemes run by different fund houses, ideally there should be different models.

In order to analyse different parameters impacting the performance of the schemes run by different fund houses, the project employes a more qualitative feature selection. Under this method project used Multiple Linear Regression technique on different schemes. Project ran the multiple linear regression individually for different schemes of a fund house and then stored the coefficients and the p-value corresponding to different variables in two matrices.

The first column of the coefficient matrix contained the name of different schemes for the given fund house. The following columns contained the coefficients corresponding to different variables. The last column of the matrix contained the Adjusted R-square for the given regression. The matrix of the p-value is similar to the coefficient matrix except the fact that the coefficient value is being replaced by the p-value for each variable. This procedure is repeated for different fund houses. On applying the conditional formatting in MS excel, it was observed that some variables were very significant in most of the regression while some of the variables were not significant in most of the regression model.

MODELLING

Before initial modelling, various plots of output(return) vs input (age of the fund, nav_week_high etc.) were looked at. Various transformations such as log(x) and $[\log(x)]^2$ were applied on different independent variables such as age of the fund, NAV_week_high, NAV_week_low etc. However, best R^2 value for linear regression obtained was 0.015. This result, as well as the plots suggested that there was a high degree of non-linearity in the data. Random forests were used for modelling. With depth = 7 and trees = 500; R^2 value of 0.13 was achieved initially. After many iterations, best R^2 value for random forest was achieved to be 0.255 at depth = 4 and trees = 400.

Number of Trees	Depth	R^2 Value
100	3	0.18
500	7	0.13
800	4	0.238
500	4	0.244
600	4	0.22
400	**4**	**0.255**
300	4	0.225
350	4	0.224

List of Parameters for Random Forest

XGBoost

XGBoost is an open-source software library which provides the gradient boosting framework. Gradient boosting is a machine learning technique used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods , and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

XGBoost was used with a learning rate of 0.1 initially. R^2 thus obtained was 0.17. Upon changing the learning rate to 0.5, R^2 value had increased to 0.25.

Linear booster was also tried but its R^2 value was very poor ~0.1. To get the best parameters, extensive parametric search was done using MLR package in R. Learning rate was fixed at eta = 0.3. The best parameters were obtained as follows:

booster=gbtree; max_depth=9; min_child_weight=9.68; subsample=0.525; colsample_bytree=0.935

R^2 value was obtained to be 0.53