EDA on Mutual Funds

Gourav K Verma

Data Science, Bellevue University

DSC530-T302 Data Exploration and Analysis

Professor Matthew Metzger

Feb. 29, 2020

Abstract

A mutual fund is a company that pools money from many investors and invests the money in securities such as stocks, bonds, and short-term debt. The combined holdings of the mutual fund are known as its portfolio. Investors buy shares in mutual funds. Each share represents an investor's part ownership in the fund and the income it generates. There are many parameters and points attached to a Mutual Funds. Out of these I will be putting my focus on important factors tied to a high performing Mutual funds.

Keywords: Exploratory Data Analysis, Data Science, Probability, Modeling, Estimation, Regression, Expense ratio, Morning star ratings, Fund Return, net asset.

Contents of Dataset

The file contains 25,265 Mutual Funds with general aspects (as Total Net Assets, Management Company and size), portfolio indicators (as cash, stocks, bonds, and sectors), returns (as year-to-date, 2018-10) and financial ratios (as price/earnings, Treynor and Sharpe ratios, alpha, and beta). Looking into data it appears data was captured around Feb-2019.

Data has been scraped from the publicly available website https://finance.yahoo.com

Important Variables

- **net_assets:** It represents the total of all dollars invested in all share classes of the fund. Do not confuse it with Net Asset Value (NAV, per share/unit price of the fund)
- **morningstar_rating:** The Morningstar Rating is a measure of a fund's risk-adjusted return, relative to similar funds. Funds are rated from 1 to 5 stars, with the best performers receiving 5 stars and the worst performers receiving a single star
- **net_annual_expense_ratio_fund:** The expense ratio is the annual fee that all funds charge their shareholders
- **investment:** Classification of funds based on both the size of the companies invested in and the growth prospects of the invested stocks.
- **size**: Size of fund. Large, Medium, Small
- **fund_return_2018:** profit generated in year 2018
- **fund_return_10years:** profit generated in past 10 year

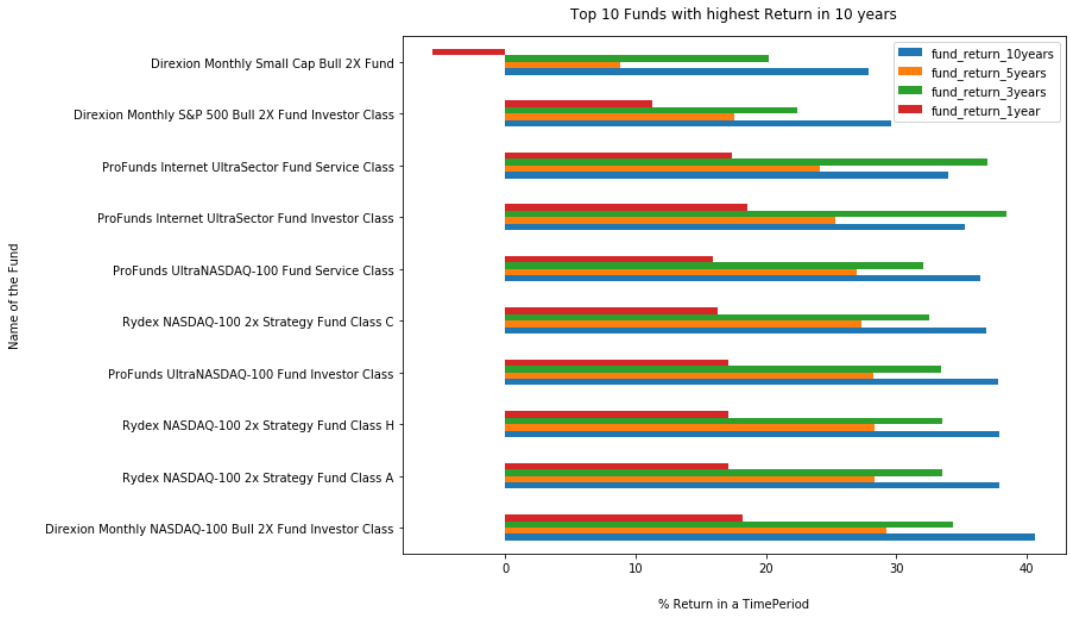
Exploratory Data Analysis Steps

Cleaning:

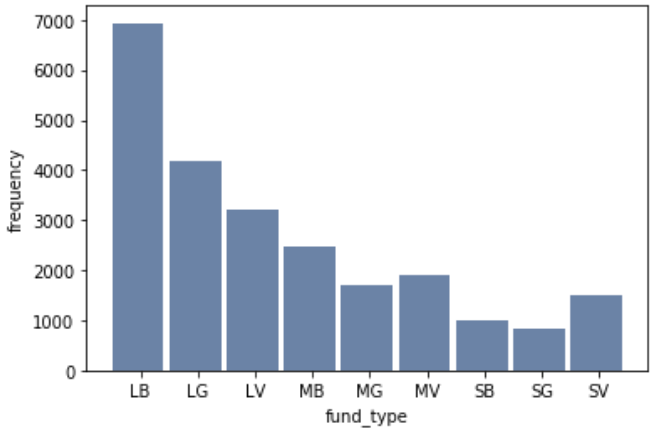
- In Mutual Fund file there were 25308 rows and 125 columns.
- Identified the important variables and created new dataframe with selected 46 variables.
- In the new dataframe there were 40 columns which had missing values. But none of them were had more than 10% missing values.
- If I had removed Null values, I would have left with approx. 4k rows, which was pretty low. So, I continued without removing null values.
- Divided net asset with 1 million for easier understanding.
- Column name ‘size’ was giving issue in some functions, so renamed it to f_size
- Created new column ‘ftype’ by concatenating columns ‘f_size’ and ‘investment’.

Outcomes

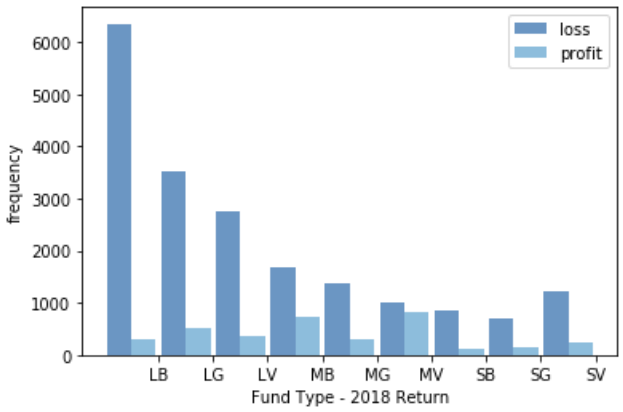
- Below plot shows funds gave highest return in 10 years and their corresponding 5yr, 3yr and 1yr returns.

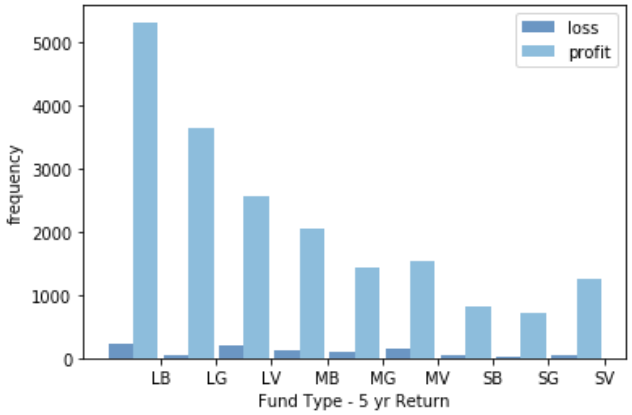


- Histogram of Fund type. Majority of funds were Large Blend and Small Growth were in minority.

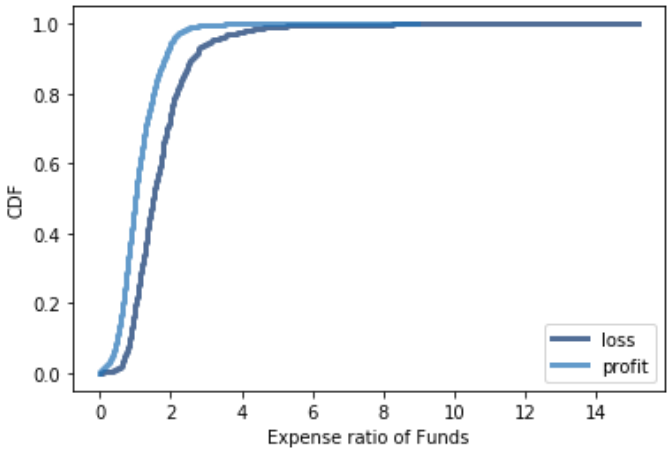


- Histogram of comparison of profit/loss based on fund type for 2018 and tenure of 5years. It clearly shows that most of the fund's returns converted into profit if the investment went for past 5 years instead of just a year. The more time you invest in Large Blend funds, more chances of having the profit.





- Below CDF graph shows that funds which had profit in 5 year tenure, had lower expense ratio.



- Prepared **model** for 10 year return based on morning star rating, net asset and price to earnings ratio.

  R-squared of the model was 22%

  formula = 'fund_return_10years ~ morningstar_rating + net_assets + price_earnings'

  As per our model there will be 9.3% return in 10year on a 200 million dollar fund, whose morning star rating is 4 and price to earnings ratio is 17.0

Dep. Variable:	fund_return_10years	**R-squared:**	0.220
Model:	OLS	**Adj. R-squared:**	0.220
Method:	Least Squares	**F-statistic:**	2357.
Date:	Sat, 29 Feb 2020	**Prob (F-statistic):**	0.00
Time:	15:16:11	**Log-Likelihood:**	-79559.
No. Observations:	25073	**AIC:**	1.591e+05
Df Residuals:	25069	**BIC:**	1.592e+05
Df Model:	3		
Covariance Type:	nonrobust	**Omnibus:**	1846.441
		Prob(Omnibus):	0.000
		Skew:	-0.512
		Kurtosis:	4.562

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.4120	0.099	-4.152	0.000	-0.607	-0.218
morningstar_rating	1.3287	0.029	46.038	0.000	1.272	1.385
net_assets	2.807e-05	2.02e-06	13.908	0.000	2.41e-05	3.2e-05
price_earnings	0.2584	0.004	63.176	0.000	0.250	0.266

Durbin-Watson:	1.294
Jarque-Bera (JB):	3644.826
Prob(JB):	0.00
Cond. No.	5.17e+04

References

Downey, A. B. (2014). *Think stats: exploratory data analysis*. " O'Reilly Media, Inc.".