**Investment Management Using Robo-Advisors**

**Catered to Individual Needs**

Amarnath Reddy[a], Ashmit P Bhardwaj[a], Saketh Varma[a]

[a] Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, NY 14260,  
USA

**I. Abstract**

The overall purpose of this project is to transform the mode of investment management using predictive algorithms, unveiling closely guarded secrets in the financial world. Recognizing the variety of personal financial needs and the absence of a deterministic path to success, our project attempts to provide access to various risk assessment and investment recommendations. The project tackles the lower signal-to-noise ratio found in stock markets in an era where the noise of financial markets can be challenging even to the most experienced investors. This strategic advantage is employed to create customized portfolios. Drawing inspiration from Forbes’ recognition of Data Scientists as the new financial advisors, we envision using predictive algorithms for market trend analysis. The strategies used will be uniquely designed to suit user-specific needs, addressing their financial targets and requirements. It is done by building investment piggybanks, one for each need, where every bank has their own risk tolerance and capital, in turn producing a need-specific portfolio based on that financial goal.

The models dynamically allocate assets according to market trends and keep tracking it against the base market (S&P500 Index). Decision Trees help us define the risk tolerance, neural networks like LSTM recommend efficient portfolio diversification while regression keeps track of predictive market behavior. In our research findings, we deduced that there is a significant roadblock between traditional financial strategies invented in the eighties, and the up-and-coming world of machine learning. The research aims to close that gap by introducing robot-advisors-based decision making, while also making sure that our models get evaluated and validated by the same proven financial methodologies as Sharpe Ratio, P-E Ratio, K-score etc.

Central to our approach are two datasets. The "Risk Appetite Dataset" profiles individual investors, highlighting their demographic and financial specifics such as age, income, gender, marital status, dependents, risk tolerance, investment maturity, potential returns, and interest rates. This dataset is instrumental in tailoring investment strategies to each investor's unique financial goals and risk preferences. The "ETF Prices Dataset" offers exhaustive historical data on ETFs, covering aspects like daily price fluctuations, trading volumes, and adjusted closing prices. This market data is key to analyzing trends and guiding asset allocation in investment portfolios.

To conclude, the project provides a platform to anyone who lacks access to traditional financial advisors or large institutions such as hedge funds and trust funds to reach financial freedom on their own terms. As we are not financial advisors, this research is purely based on generic predictive modeling algorithms and shall not be considered financial advice without context or verification by a professional, who provides their insights into stocks, inflation, and volatility. In summary, by harnessing the power of Data Science, we are determined to encourage people on their quest for financial success, one data-driven investment at a time.

**II. Literature Review**

**2.1 Predictive Modeling in Finance**

This section discusses research on deep learning in financial markets, focusing on capturing complex structures in financial data. The papers explored include studies on stock market forecasting using deep learning networks, efficacy of deep learning in stock prediction, and the integration of autoencoders and LSTM for financial time series forecasting.

This [1] paper highlights the universal patterns in price formation in financial markets by utilizing deep learning methodologies. Their study illuminates how these machine learning techniques can capture intricate structures in financial data, providing a fresh perspective on the dynamics of market prices and its potential in capturing universal features in price formation.

This [2] paper explores the potential of deep learning networks in the domain of stock market forecasting. The study delves into the methodologies employed, different data representations suitable for financial time series, and presents real-world case studies to demonstrate the efficacy of these deep learning models in predicting stock market movements. The paper serves as a comprehensive guide for integrating advanced neural networks in financial analysis.

The authors in [3] conducted a detailed analysis of the efficacy of deep learning networks in stock market prediction, focusing on high-frequency intraday stock returns. The study explored three unsupervised feature extraction methods and how they influenced the network's predictive capabilities. Empirical findings indicated that while deep neural networks enhance prediction performance by extracting from autoregressive model residuals [10], the reverse does not hold true. The research also revealed improved covariance estimation in market structure analysis using predictive networks.

Bao et al. [4] presented a deep learning framework tailored for financial time series forecasting. By integrating stacked autoencoders [8] and long-short term memory (LSTM) [6] networks, the study aimed to capture intricate patterns in financial data and improve the precision of financial predictions. This approach sought to bridge the gap between traditional time series models and the potential advantages of deep learning methodologies in financial analytics.

[5] utilizes long short-term memory (LSTM) neural networks to predict the next day’s closing price of the S&P 500 index using a mix of market data, macroeconomic data, and technical indicators. Comparative analysis reveals that the single layer LSTM model outperforms multilayer LSTM models in predicting stock market behavior based on standard metrics.

The study [7] introduces the Probabilistic Sharpe Ratio (PSR), an uncertainty-adjusted metric for evaluating investment skill, which balances record length with statistical characteristics and allows for optimal portfolio construction under non-Normal returns.

Conclusively, [9] examines and compares various predictive algorithms—Backpropagation, Support Vector Machines (SVM), Long Short-Term Memory (LSTM), and Kalman Filter—in forecasting stock market values. The comparison is based on accuracy, variation, time required for epochs, and reliability, as assessed by the T-test hypothesis test.

**2.2 Machine Learning Techniques for Portfolio Diversification**

This section reviews studies applying machine learning to stock ranking and portfolio diversification. It includes works on technical features-based random forest models, learn-to-rank algorithms for stock selection, and comparative studies of various machine learning techniques in predicting stock market behavior.

Christian et al [11] derived a stock ranking by applying a technical features-based random forest model [20] on an international dataset of liquid stocks. Rather than predicted return, the ranking is based on outperformance probability. By applying a decile split, he found that long and short portfolios achieve Sharpe ratios of up to 1.95 and a highly significant yearly six-factor alpha of up to 21.79%.

Similarly, Alsulmi et al [12] also ranked the stocks using learn-to-rank algorithms [21] (LTR) to rank the Saudi stocks by creating several simulated long-term investment portfolios with an investment. This study highlights the viability of leveraging computational intelligence for more effective and profitable stock selection strategies.

Saha et al [13] worked on stock ranking using a graph-based strategy, which combines relational information about stocks for stock ranking prediction, addresses the shortcomings of conventional stock prediction techniques. They suggest a brand-new evaluation metric for stocks called normalized rank biased overlap for top-k stocks that is designed to account for the importance of top-k stock predictions in investing choices. The paper also emphasizes how list-wise loss functions and node embedding methods like Node2Vec [28] can improve stock ranking prediction performance, presenting potential advantages in investment strategies over conventional stock-based approaches.

Qiang et al [14] used learning-to-rank algorithms to create trading strategies based on a group of equities' relative performance, with an emphasis on investor sentiment. The study develops stock selection guidelines that call for long positions in the top 25% of ranked equities and short positions in the bottom 25%, as decided by learning-to-rank algorithms, by using sentiment shock and trend indicators from earlier studies. A decade's worth of market and news sentiment data are used to assess the performance of two algorithms, ListNet and RankNet, which are specifically used in the stock selection process. Using back testing from 2006 to 2014, the research shows that these portfolio strategies outperform the S&P 500 index, the hedge fund industry average performance (HFRIEMN), and some sentiment-based approaches lacking learning-to-rank algorithms during the same period. The success of learning-to-rank algorithms in the context of sentiment-based stock trading methods is highlighted by this study.

Robert et al [15] addressed the stock selection problem using a novel method called Prototype Ranking (PR), which takes on the problems brought on by the size of actual stock data. To predict stock rankings, PR uses a modified form of competitive learning, with an emphasis on picking out the best-performing stocks from a group of average ones. High-performing stocks are typically in the minority in noisy stock samples, where this method is specifically created to excel. PR undergoes evaluation through a trading simulation using actual stock data to judge its performance. In this simulation, a portfolio is built each week using the stocks with the highest predicted ranks.

Huang et al [16] compare machine learning-based linear models to traditional regression-based models for stock scoring in investment and financing. It illustrates that the genetic-based approach surpasses both conventional regression techniques and benchmark models, employing Genetic Algorithms [29] for optimization of parameters and input variable selection. This suggests that GA-enhanced machine learning techniques offer an exciting opportunity for stock selection, eventually advancing the field of machine learning in finance and providing a compelling alternative to conventional regression-based approaches.

Hota et al [17] worked on predicting stock prices using models including Decision Tress [32], Support Vector Regression [33], Random Forest [20], and Artificial Neural Network [34]. According to their observations, Random Forest outperforms the other models, and it is indicated as the best tool for real-time implementation in stock price prediction.

Matthias et al [18] forecasted the distribution of emerging market stock returns and examined various machine learning methods. They specifically examined the predictive performance of nine methods, including neural networks with one to five layers and tree-based models such as gradient-boosted regression trees and random forest. Traditional linear models included ordinary least squares regression and elastic net as examples. They also looked at the performance of an ensemble that included two tree-based models and an ensemble of neural networks, five different neural networks, and approaches that can manage non-linearities and interactions.

Jour et al [19] graph-based approach to stock market analysis is used in this paper, which indicates intricate patterns within the interconnected structure of individual stocks and market indices. From five viewpoints, this survey investigates new areas of the stock market graph formulation, graph filtering, clustering, stock movement prediction, and portfolio optimization. It gives an in-depth review of key techniques and algorithms, puts light on the evolving landscape of graph-based methodologies in stock market analysis, and states suggestions.

This [30] study reviews existing research on diversified stock portfolios and examines how the optimal number of stocks has evolved over time and in different market conditions, including the Global Financial Crisis and the COVID-19 pandemic. It concludes that there is no one-size-fits-all answer for the ideal number of stocks in a diversified portfolio. Contemporary trends suggest that diversified portfolios now require more stocks, emerging markets need fewer stocks, high stock-market correlations reduce the need for many stocks for individual investors, and machine learning can enhance investment decision-making. These findings are valuable for investors and lay the groundwork for future research.

* 1. **Monte Carlo Simulation in Risk Assessment**

This part examines the use of Monte Carlo simulation in risk management across different sectors. It covers studies on risk assessment in construction projects, economic risk analysis of renewable energy technologies, and financial risk assessment in highway infrastructure projects.

Risk management is a fundamental aspect of project execution across various sectors, necessitating the identification and analysis of risks [22]. Precise risk identification and in-depth analysis contribute significantly to cost and time savings, as well as improved project quality. The Monte Carlo simulation method has been widely adopted in research within the risk management domain to enhance risk identification and analysis techniques [23,24]. In the subsequent sections, we will examine several pertinent studies in this area.

Albogamy and Dawood [25] endeavored to formulate an effective risk assessment approach for client-related risk factors that exert substantial influence on the success of construction projects spanning the preliminary design to construction and operation phases. Their approach incorporated the Monte Carlo simulation method, leveraging probability and random variables to manage risks associated with escalating costs and project scheduling delays. Their primary objective was to establish a client-centric risk management model. The proposed methodology integrated the Analytical Hierarchy Process (AHP) with Monte Carlo simulation, aligning it with a developed risk plan, and substantiated its utility through a comprehensive case study.

Arnold et al. [26] employed Monte Carlo simulation to assess the risks associated with investments in decentralized renewable energy technologies (RETs) within a life cycle framework. The central predicament they sought to address pertained to the tepid interest of investors and institutional lenders in decentralized RETs due to economic barriers, including prohibitive initial capital expenses, transaction costs, and multifarious project-related risks. They underscored the need for advanced risk management tools to mitigate transaction costs and financial risks and proposed that combining financial analysis with Monte Carlo simulation could lead to more robust conceptual designs for investment projects, enhancing their financial viability and risk profiles. This approach, they suggested, could foster increased investments in technology projects, thereby promoting decentralized renewable energy initiatives.

Kumar et al. [27] harnessed the power of Monte Carlo simulation to establish a standardized risk modeling method for assessing financial risks associated with highway infrastructure projects, particularly in relation to variables like traffic flow and project cost. Their methodology entailed employing the Net Present Value (NPV) model in conjunction with Monte Carlo simulation to scrutinize the probability distributions of input variables introducing uncertainty into NPV calculations. The researchers contended that this approach serves as a valuable decision-making tool, aiding in determining the feasibility and profitability of a project. Additionally, it assists in identifying the types of uncertainty with the most pronounced impact on a project's financial performance, enabling governmental agencies and officials to devise measures to mitigate these uncertainties' effects on projects.

Ahmadi-Javid et al. [31] authored a scholarly article that advocates an imperative shift toward a holistic project portfolio risk management approach, featuring an initiative-taking, mathematically optimized strategy. It introduces an innovative methodology encompassing a spectrum of considerations, including cost, budget, preferences, risk-event probabilities, and nuanced assessment of risk occurrence and impact dependencies, thereby extending its applicability beyond portfolio management to enhance project-specific risk mitigation.

* 1. **Research Gaps**

As highlighted by Heaton et al. [34], the distinctive feature in the application of machine learning (ML) tools in finance, compared to other scientific domains, lies in the fact that finance focuses not on replicating tasks humans excel at. Unlike tasks such as image recognition or responding to verbal requests, humans lack an innate ability to, for instance, identify stocks likely to perform well in the future. Consequently, the value of ML tools in the financial context is sought in different avenues. Specifically, their potency is particularly pronounced in selection problems because, fundamentally, they offer the most efficient and expedient means of computing functions that map data, encompassing returns, prices, economic indicators, accounting figures, and the like.

In [36], the authors point out that the world of financial machine learning is a closed one, as companies with extensive models, data, and assets to play with always succeed. Sharing their strategies would mean a loss in value for their firm. Hence, they are only accessible to high-net-worth individuals.

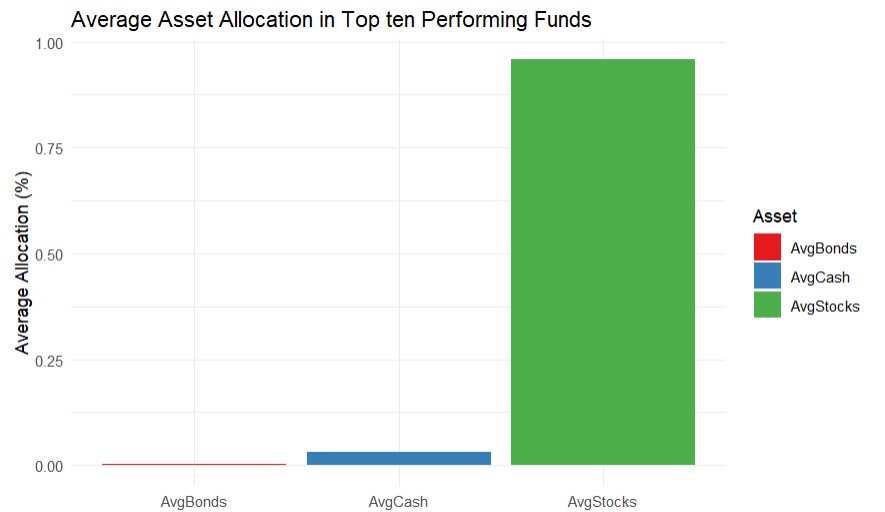
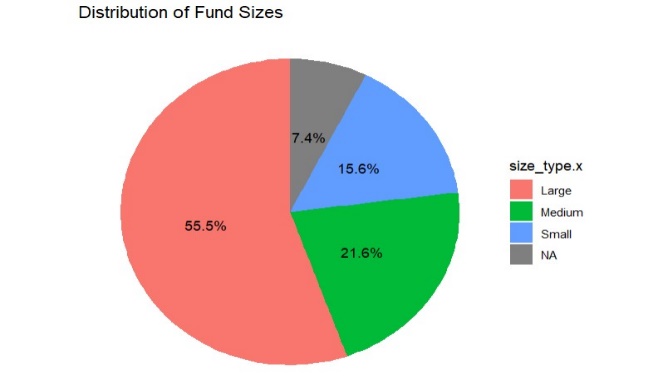
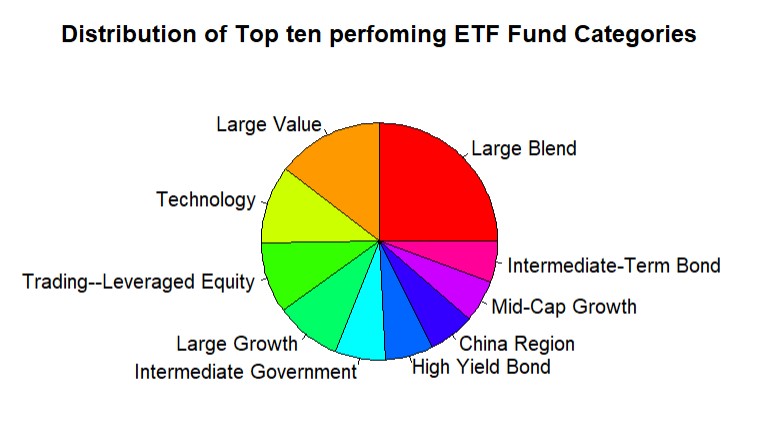
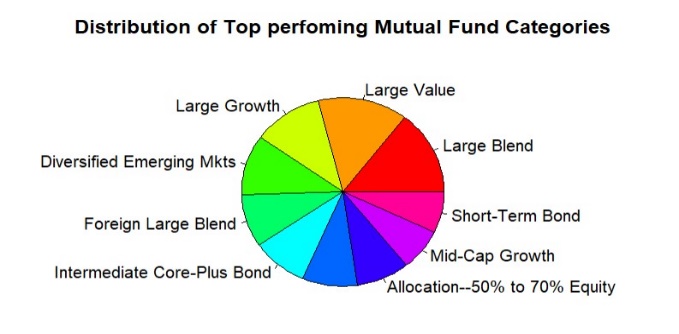
Also, in our survey we found that none of the researchers combined various techniques to create an end-to-end model where they catered to individual needs while also diving deep into stock picking and portfolio diversification. Our aim is to combine these techniques and form a service available to all, which provides the platform for financial independence, even to those whose voices never reach the ever-glowing kingdom of hedge funds but whose financial goals are just as important, if not more, for the advancement of society at large.

**III. Exploratory Data Analysis**

**3.1 Mutual Funds and ETFs**

We considered Mutual Funds data from 2000 to 2020. The data was scraped from Yahoo Finance which can also be found on Kaggle. We start by showing the basic distribution of how ETFs and mutual funds invest in various kinds of stocks.

As seen clearly in Fig. 1, the largest distribution is held by large cap funds. This makes sense because these are the companies that have shown themselves to be the most stable revenue-generating businesses in the past.



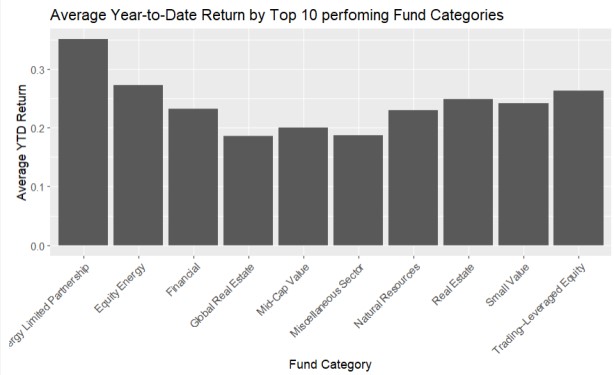
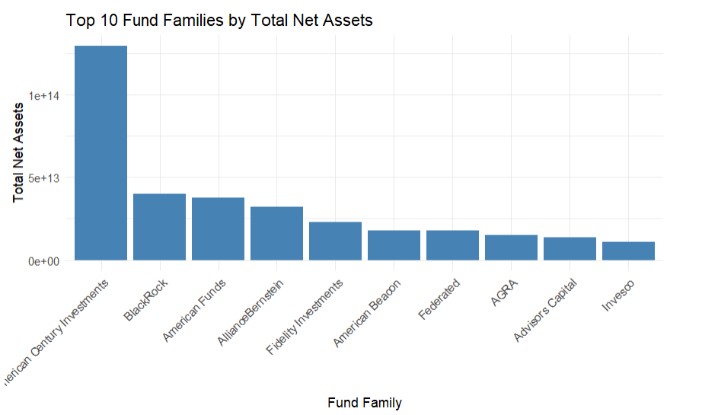
# Figure 1 – (top) Distribution of top performing categories, (bottom left) Fund sizes pie chart, (bottom right) Allocation Distribution

When we observe the top ten funds based on size, it is much clearer how important large cap companies are to the functioning and analysis of these funds. They hold more than half of all the assets of these holding companies.

Next, with a bar graph we analyzed how investments are distributed across assets. Bonds are inflation-matching funds, which would return to us a steady 3-5% growth rate. Cash is held just in case some new opportunities arise and the decision firm does not want to disturb their existing holdings to cash in on the growth. Hence, a fair amount is kept aside in cash. Although it may hurt in the short term as inflation eats upon cash quickly, over the years when this money is invested in a quick growth market, the overall returns average out to return a better investment than initially thought of. But, as always, most of the investments are in stocks as these would be our key performers, returning the most value for money. Depending on the stocks we choose, the CAGR can range from 5-20% if properly diversified and maintained.

Then, we visualize the fund families and how much assets they manage in Fig. 2. As expected, American Century Investments top the charts followed by BlackRock and American Funds. Keep in mind that these funds are for publicly available retail investors and the numbers would look staringly different if private equity came into picture, where BlackRock and Vanguard control almost 90% of the market.

To check which funds have performed the best in the current dataset, we compute their average return weighted by fund holdings and observe that Energy Limited is the top performing fund in this regard. But, before we put all our eggs in the best basket, we need to remember that, just like in any model, historical performance does not guarantee future outcomes.

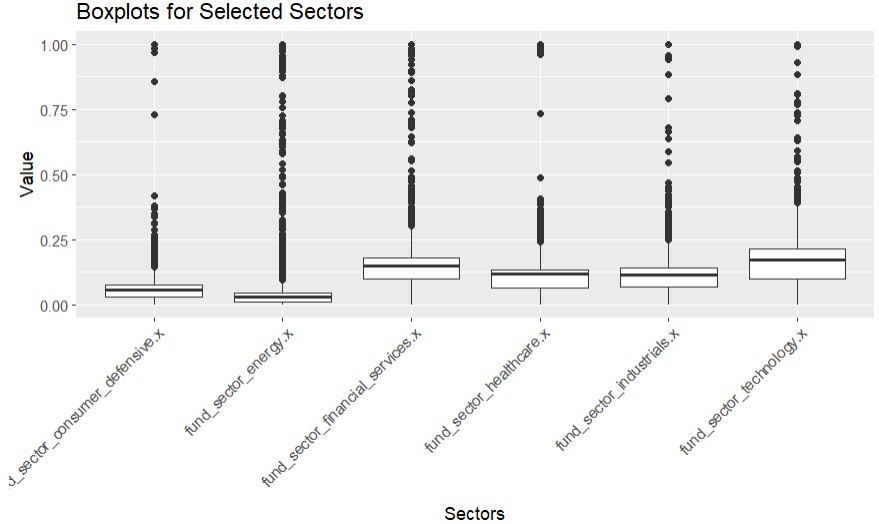
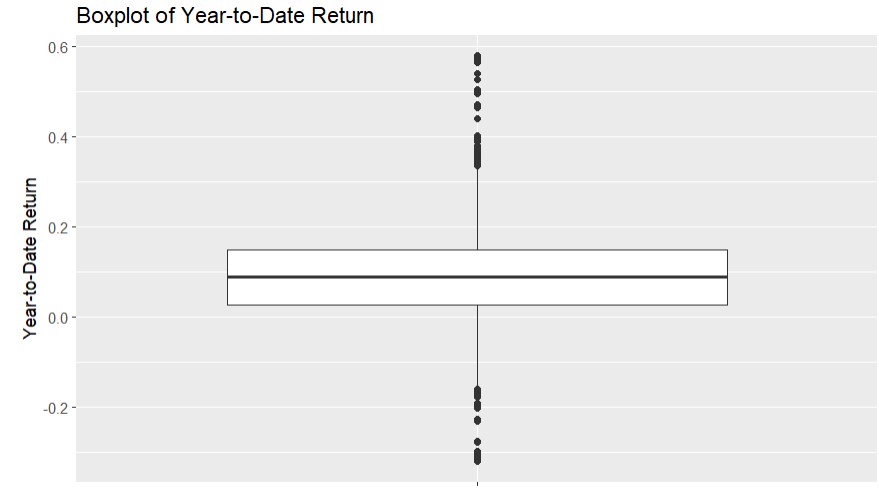
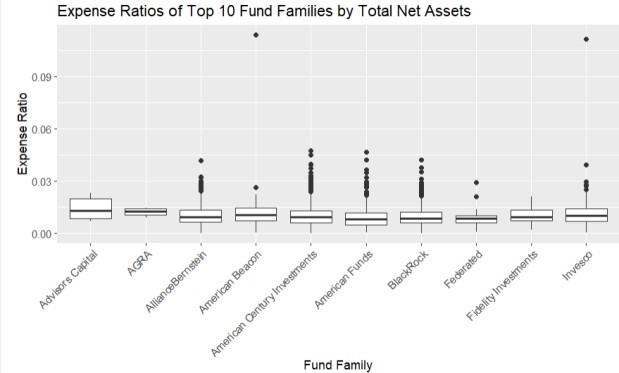


# Figure 2 - (left) Top 10 funds families by Total Assets, (right) Return by Fund categories.

Other than returns, the most important facet to keep in mind is the expense ratio for these funds [Figure 3]. As the expense ratio acts like a fee for our invested money that we never see, despite of whether we gain or lose money, it is important to invest in a low expense fund like ETF where we see every dollar we earn back. Most of the funds have an expense ratio centered around 0.015 which is negligible unless we are investing a considerable sum (say million dollars). Safe to say that we can ignore this statistic for now, even while building a model, but we will make sure to review it at the end to make sure we are not losing money where we should not.

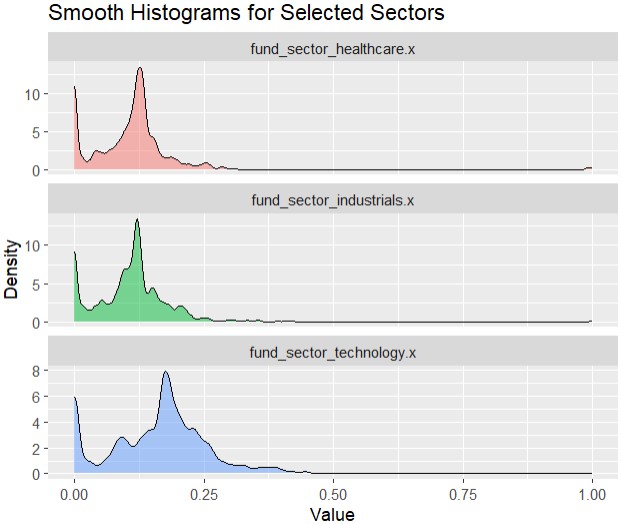
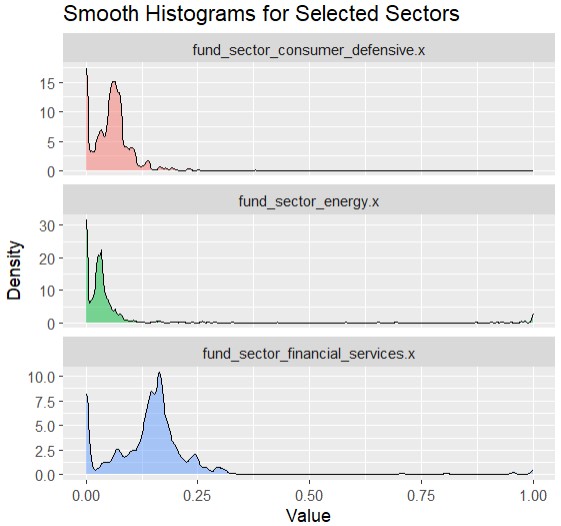
The boxplot for YTD returns shows that most of the funds are barely beating inflation, hence, the need for a financial advisor is even more necessary to generate the best returns. Only outliers are producing above 4% of annualized returns consistently so our research will be important in its selection.

Finally, we observe the boxplot for the sector and their values and see that finance and technology sectors have a slight edge over the others. This makes sense as the companies that we know about and hear about everyday are financial and technology companies, so its stock is considered a growing stock. They can also easily serve a larger market (think Facebook coming to India) with just a few tweaks and the scope of expansion for these firms is far more than healthcare or defense sectors, which cater to a niche and expansion for them might cost in terms of getting security clearance or running trials to clear bureaucratic barriers.



# Figure 3 – (top left) Expense Ratio by Net assets, (Top right) Boxplot Returns YTD, (bottom) Boxplots for selected sectors.

We can also observe the same trend when we plot the smooth histograms in Figure 4 of each major sector. It clearly shows that there are more funds on the higher end of the 0-0.25 range, with some creeping up even higher. This means that we should pay attention to these funds in our model, as the actual financial experts are doing.



# Figure 4 – Distribution of Sector Percentages held in each fund.

When we plot alpha (see glossary) for 5 years differentiated on fund families, we find that the more we invest in healthcare, the more odds we must beat the market. The same cannot be said for the industrial or defensive sector however, where we see a slight decline in the regression line. For others, it remains constant. We need to keep in mind that this is based off only 5 years of alpha values and can be skewed by market trends or industry shifts, so we need to include more data in our models for final decisions.

A graph of different colored lines

Description automatically generated with medium confidence

# Figure 5 – Regplot Alpha by Sectors

Regarding the regression plot for standard deviation vs fund families, we observe that industrial sector has a remarkably high deviation and hence, might be a risky investment. Banking, as discussed earlier, seems like a safe bet as it stays constant throughout, both in terms of alpha and standard deviation. We can further break it down during our individual stocks analysis.

A graph of different colored lines

Description automatically generated with medium confidence

# Figure 6 – Regplots std dev by Sector

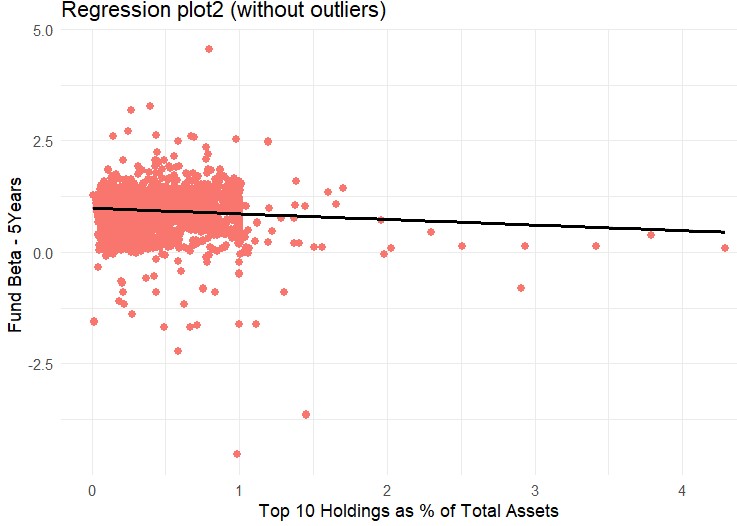
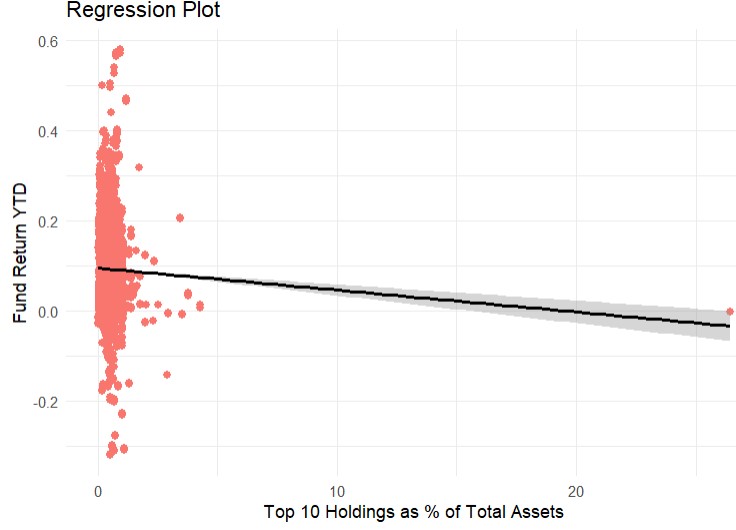
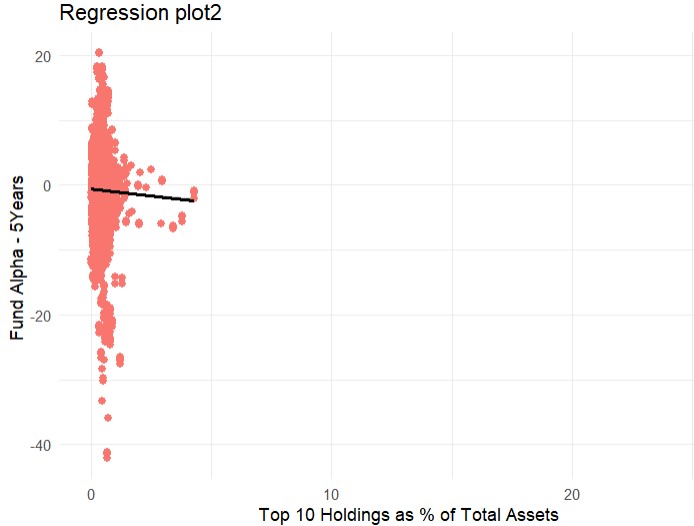
In funds return vs fund family, we see that all graphs trend upwards which is expected as inflation increases, the prices for goods and services increase as well, and the revenue and stocks see an uptick. We can see that defense sector drops by a few points which might just be a bad phase for the recorded period in that sector. If we see this trend continuing for too many years, however, we will make sure to pay extra attention to penalize our model if it adds these stocks to our portfolio.

A graph of different colored dots

Description automatically generated with medium confidence

# Figure 7 – Regression plots Returns by sector.

The next graph is important for the insights into the data and for what our model should aim. The top ten holdings as % is a ratio of total assets held by the top ten companies in that fund. The funds with lower top ten holdings % are the funds that diversified more than concentrate, and vice versa. So, the pattern observed when plotting top ten holdings against alpha, beta and returns is that the line tends to decline as the holdings increase. This means that the companies that concentrated their portfolio in their top ten stocks saw a decrease in alpha, beta and returns. What this suggests is that we need to have a diversified portfolio of more than ten stocks from each of the sectors to have a balanced roster of assets on hand, which gives us the chance to grow our money without too many risks.



**Figure 8 – (top) Alpha vs top ten holdings %, (bottom left) YTD Return vs top ten holdings %, (bottom right) Beta vs top ten holdings %**

# 3.2 Stocks

We used three thousand companies’ data from the year 1980-2020, tracking their stock prices and trade volumes for every day. Another dataset tracked company financials from 2010-2020, which included financial statements containing cash flow, EBIDTA, revenue, profit-loss etc. This was much easier to find as many APIs offer the dataset. We used Yahoo Finance API, stored it into an excel and computed the following results:

PE ratio is a crucial decision variable for stock selection. It is easy to explain and compute and makes sense intuitively. When we found the correlation between PE ratio with respect to other financial data, cash flow was the highest correlated value, with the score of 0.62. Others include revenue increase and EBIDTA.

A comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of a comparison of

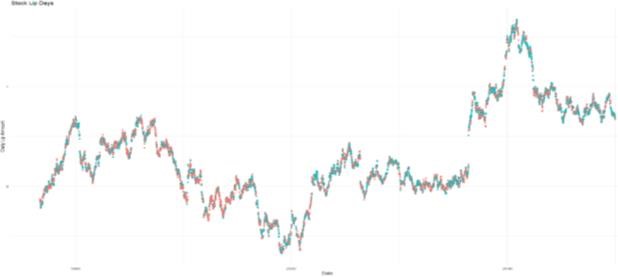
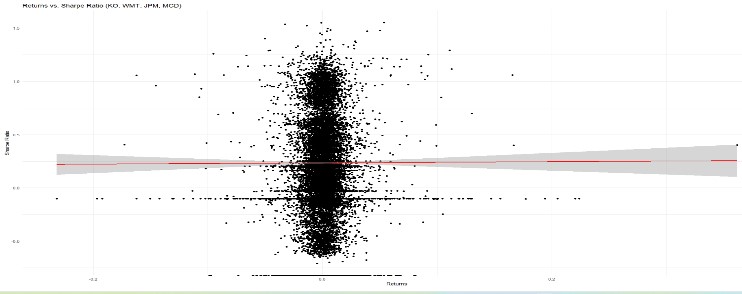
Description automatically generated

## Figure 9 – Sortino Ratio and Sharpe Ratio vs Boolean increase in day boxplot

First, we check whether the two most important technical indicators – Sortino ratio and Sharpe ratio. We split them by Boolean value of whether the stock was up that day. It shows that there is not any difference between the up and down days, showing that it is much more nuanced than that.

Next, we plot candlesticks in figure 10 top right, left and bottom left, one of the most important graphs in technical analysis. Every green indicates that it was an overall positive day, and red indicates that it was a negative day i.e., it closed lower than it opened. On y-axis, we plot the Sharpe ratio, and we observe an interesting trend in all the three stocks that we evaluated. As seen in the figure by their company logos, Amazon, Walmart, and Chase portray one behavior – whenever there are multiple reds, the Sharpe ratio will drop, and whenever there are multiple greens, the Sharpe ratio will rise. This tells us that the Sharpe ratio is a good predictor of future risk associated with the stock. If the stock rises too sharply i.e., too many good days, the Sharpe ratio falls, indicating the increase in standard deviation which is a sad thing for a stock. Any stock that can rise sharply can also fall sharply citing seasonality or just hype around that stock. Hence, it is best to stay away from these stocks.

We also observe in Fig. 10 bottom right, that there is no relationship between Sortino ratio and Sharpe ratio, which is important because our loss function will be based on maximizing these two indicators, and we cannot have a correlated loss function. So, we can use them both as our optimizer.



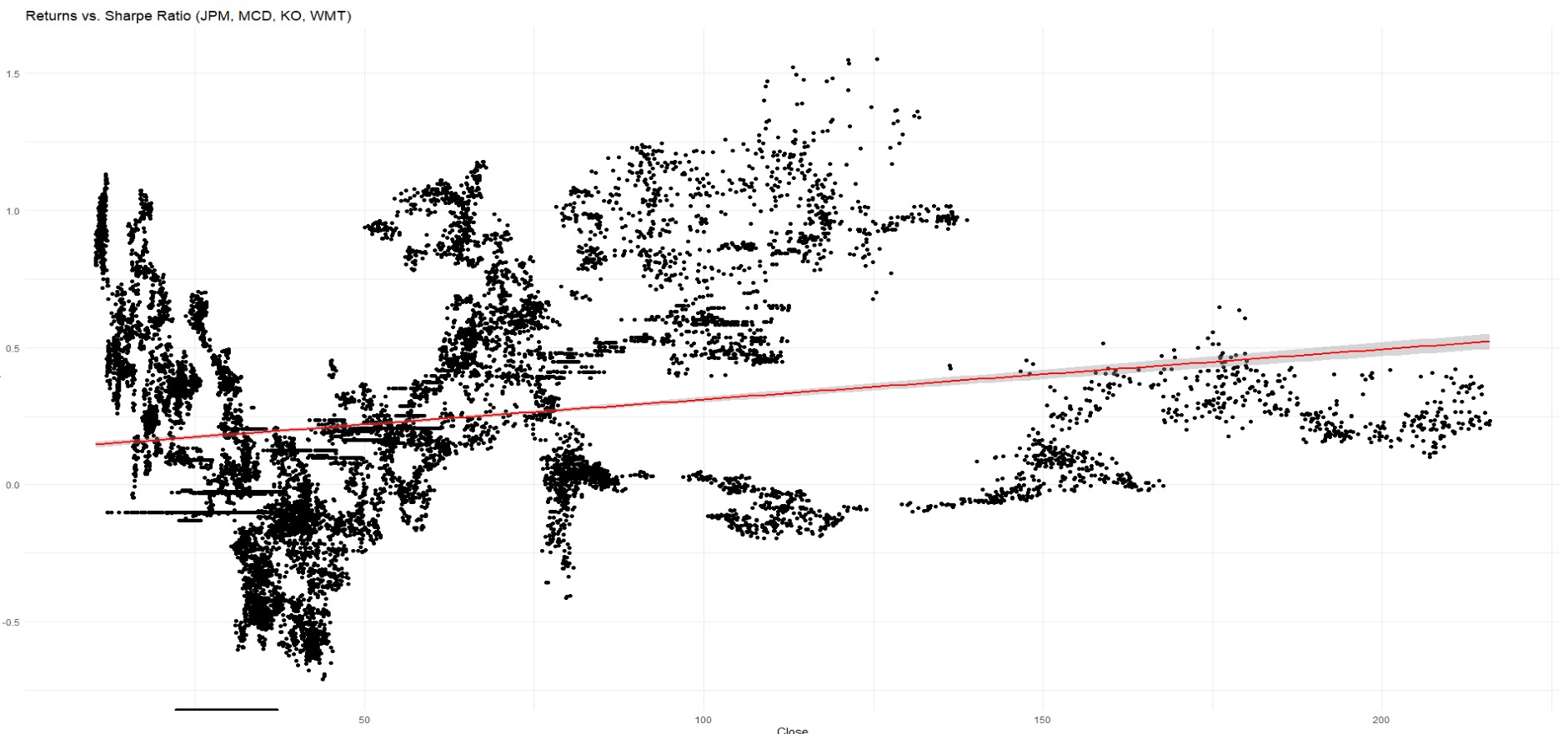
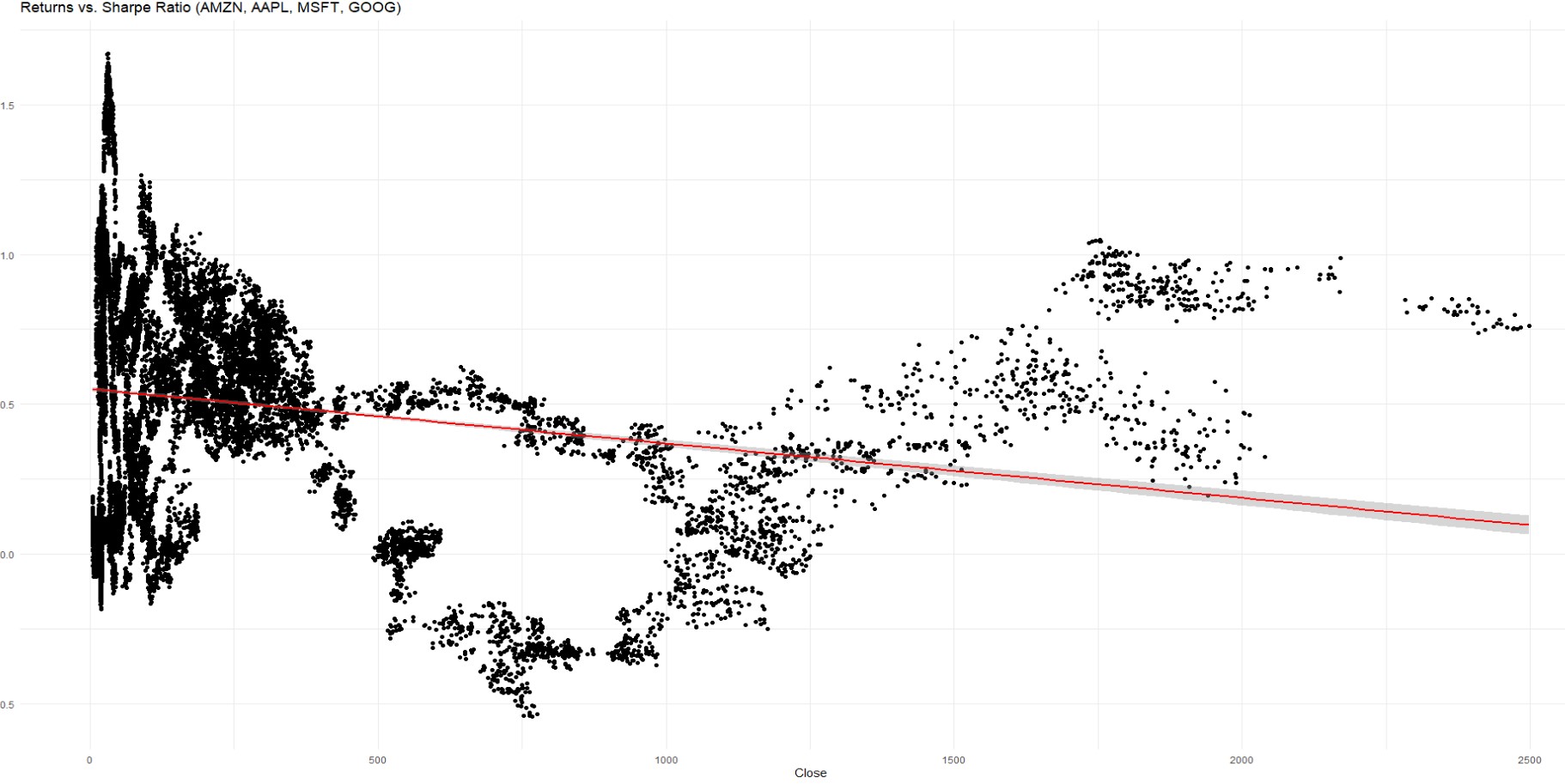
## Figure 10 (top left) Walmart candlestick vs Sharpe ratio, (top right) Amazon candlestick vs Sharpe ratio, (bottom left) Chase candlestick vs Sharpe ratio, (bottom right) Sortino ratio vs Sharpe ratio

Next, in figure 11 we plot the average PE ratio of all stocks split by their sectors. This can show us which sectors are more stable and which sectors are volatile. This can tell us if we are holding a position (buying or selling a stock), how long after we re-evaluate our stance. We see that utilities and basic materials stocks are constant with a healthy PE ratio. While healthcare is also consistent, it does have an exceptionally low PE ratio and might affect our portfolio’s overall return, even if it provides stability. Financial and technology stocks show varied behavior and although they might give the highest PE, we should make sure we have some safe bets to counter them.

A graph of lines and numbers

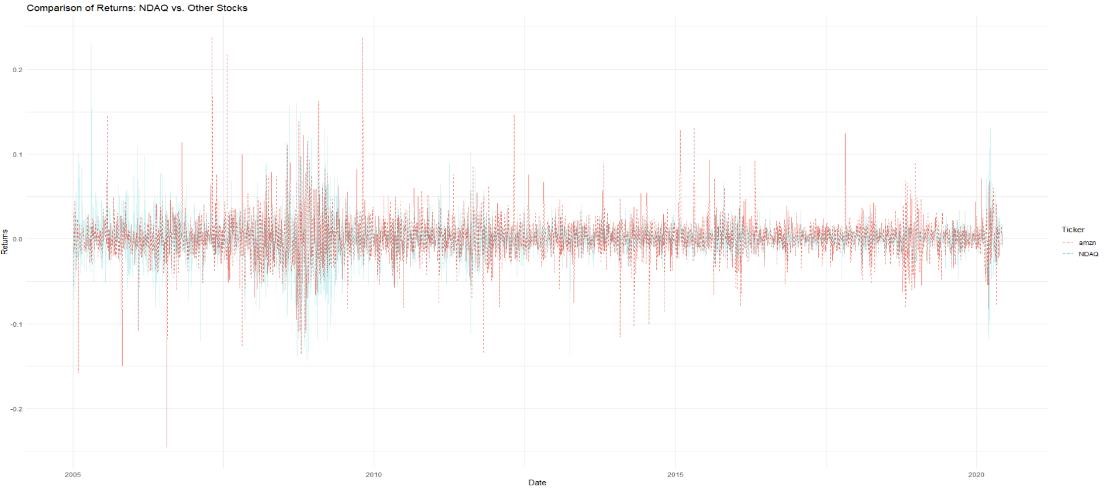
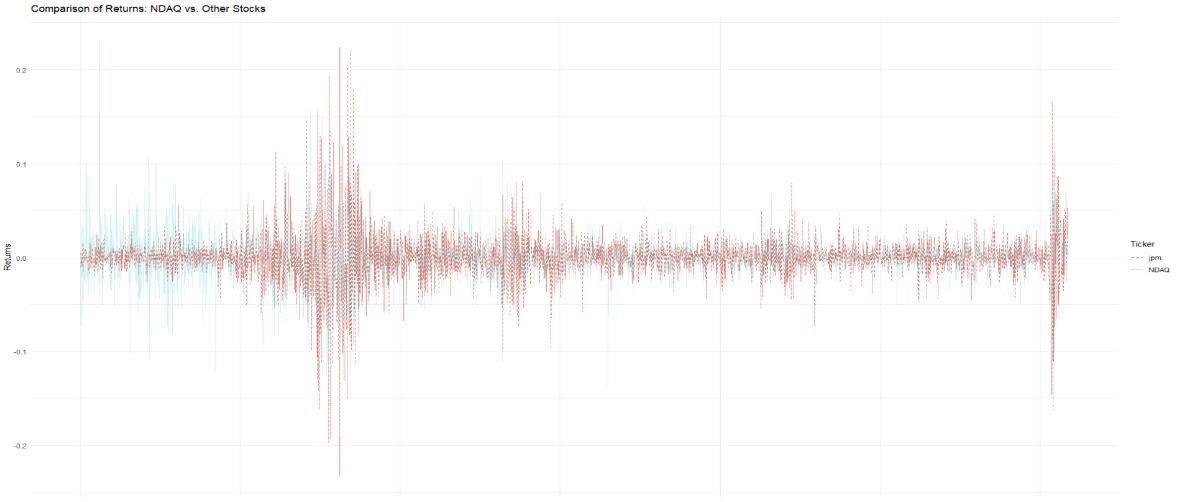
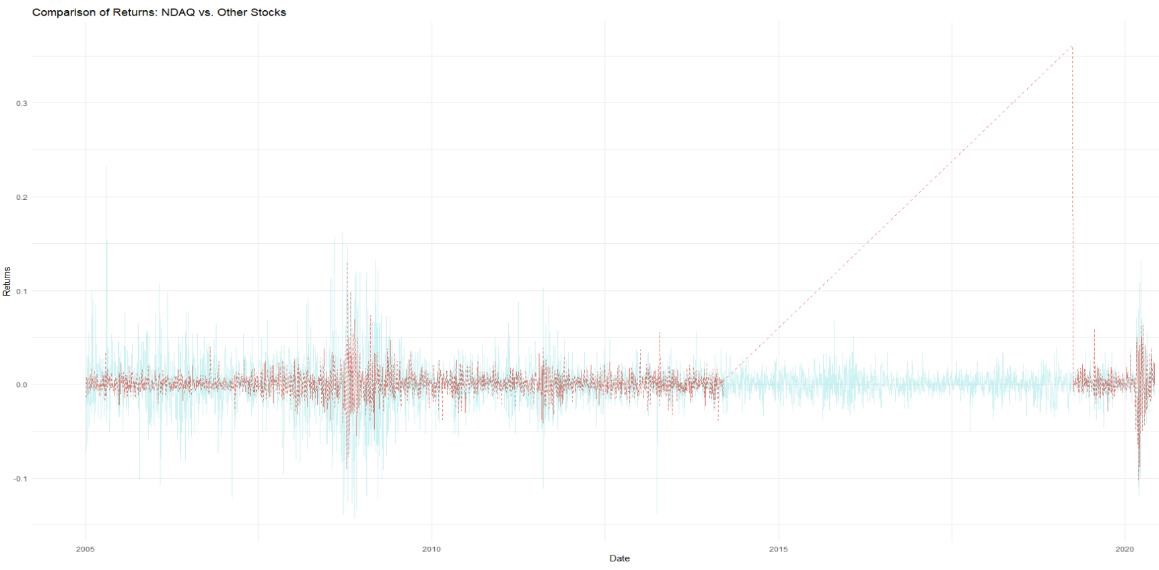
Description automatically generated with medium confidence

## Figure 11 – PE Ratio based on Sectors.



## Figure 12 – from top to bottom – “new” stocks against Sharpe ratio Replot, “old” stocks against Sharpe ratio Regplot.

In figure 12, we see two important observations – for the stocks that are in the news every day due to their recent success, the higher the price point, the worse the stock is based on Sharpe ratio. Hence, in the figure 12 top, we notice a downward trend in the regression line. On the other hand, legacy stocks that have been market leaders since the 1950s like Coca-Cola and Chase, the stocks rise as price increases. Hence, we need to observe how consistent the stock has been over the long run to determine how good of a score Sharpe ratio is. If the stock is old, Sharpe ratio has a lot of data to evaluate and gives a better judgement compared to new stocks.



## Figure 13 – from top to bottom – Coca-Cola against NASDAQ, JPM against NASDAQ, Amazon against NASDAQ

At last, in figure 13, we observe the following:

* Technology stocks do not care about the NASDAQ and are very volatile.
* The banking sector goes hand in hand with NASDAQ as the majority of NASDAQ are banking sector. Also, it relates because when the market is up, people spend and buy more, take more loans and the stocks reflect that.
* The legacy companies stay consistent despite the fluctuations in the market, barring some catastrophic events.

Hence, we think an exhaustive study of the dataset is now under the books and we have noted what data tried to tell us through various charts and statistics. We have a loss function ready at play and some models in mind to move on to the next phase.

**IV. Research Methodology**

Our research unfolds in a methodical two-step process, aiming to construct tailored investment strategies that align with individual financial goals. In the initial phase, users provide essential financial details, including their current savings, monthly investment capacity, and specific aspirations. Each distinct financial goal is referred to as a "basket," delineating the target amount and the designated time horizon for achievement.

To facilitate the estimation of investment returns needed for each basket, we employ a sophisticated computational model known as Random Forest. This model incorporates user-input data, factoring in risk tolerance and investment specifics, to generate predictions for expected returns. These predictions, in turn, facilitate the calculation of the initial investment required to attain the desired financial goal, utilizing a straightforward compound interest formula.

Transitioning to the second phase, our research leverages a comprehensive dataset encompassing financial data of S&P 500 companies, which includes detailed financial statements and year-on-year growth metrics. Adopting a classification model inspired by Warren Buffett's renowned value investing strategy, stocks undergo rigorous assessment for intrinsic worthiness, resulting in the exclusion of unsuitable stocks.

For the identified valuable stocks, we implement a Long Short-Term Memory (LSTM) network to forecast their year-on-year growth over the subsequent five years per $100 invested. The LSTM model is selected for its emphasis on predictive accuracy over complex interpretability, aligning with the forecasting requirements of our study.

Subsequently, a linear regression model comes into play, integrating past returns per $100 invested as predicted by the LSTM. This model incorporates a customized loss function, accounting for financial beta, with the primary goal of maximizing beta while penalizing deviations from the expected returns calculated in the initial phase. The outcome of this regression model is crucial in aligning the predicted returns with the anticipated returns derived from the first module.

In conclusion, the coefficients derived from the predictive regression model serve as the basis for determining stock weightages within the portfolio. Stocks with higher weightage are accorded preference in investment allocation, with proportional distribution of the principal amount. In instances where funds prove inadequate for direct stock acquisition, we employ the k-means

## Figure 14 – An overview of our approach described as a flowchart.

A diagram of a graph

Description automatically generatedalgorithm to identify a mutual fund whose characteristics closely match the desired portfolio, ensuring a nuanced and personalized investment approach. Our research methodology amalgamates advanced computational models and financial principles to construct bespoke investment strategies that are finely attuned to individual financial aspirations.

**4.1. Models Used**

Before we start explaining what individual modules do, we explain what pre-existing models we have implemented in our research. This is an exhaustive list of the models trained and evaluated, that will be referenced later at different points wherever necessary.

**4.1.1. Decision Tree**

The Decision Tree [37] is a widely used supervised learning method applicable to both regression and classification tasks. Its objective is to predict a target variable by employing straightforward decision rules derived from the dataset and its associated features. It does so by iteratively splitting the dataset in terms of the predicted variable information gain. Noteworthy advantages of this model include its interpretability and versatility in addressing problems with diverse outputs. However, a common drawback lies in the potential construction of overly intricate trees leading to overfitting.

**4.1.2. Random Forest**

A considerable collection of decision trees constitutes a random forest model [38]. This model takes the average of the predictive outcomes generated by individual trees, collectively forming a "forest." The algorithm introduces three key randomization concepts: the random selection of training data when constructing trees, the random choice of variable subsets when dividing nodes, and the consideration of only a subset of all variables for splitting each node in every foundational decision tree. Throughout the training phase of a random forest, each basic tree learns from a randomly chosen sample of the dataset.

**4.1.3. Adaboost**

The technique known as Boosting [39] aims to enhance the predictive power of weak learners, transforming them into a robust entity. AdaBoost [40], a specific form of Boosting, operates as an ensemble model, refining the predictions of each learning technique. The objective of boosting is to sequentially train weak learners, adjusting their previous predictions. This model acts as a meta-predictor, initiating the process by fitting a model on the initial dataset and subsequently fitting additional copies of it on the same dataset. Throughout the training process, the weights of samples are adjusted based on the current prediction error, thereby directing the subsequent model's focus towards challenging instances.

**4.1.4. XGBoost**

XGBoost [41] is a contemporary ensemble model rooted in decision trees, utilizing Boosting principles for weak learners. Introduced for superior performance and speed relative to other tree-based models, XGBoost offers notable advantages. These include regularization to counter overfitting, built-in cross-validation capabilities, adept handling of missing data, awareness of interactions between variables, parallelized tree construction, and effective tree pruning.

**4.1.5. Long Short-Term Memory**

Long Short-Term Memory [42] (LSTM) is a type of recurrent neural network [43] (RNN) architecture that is designed to address the vanishing gradient problem, which can occur in traditional RNNs. LSTMs are particularly adept at capturing and remembering long-term dependencies in sequential data.

In the context of sequential data, such as time series or natural language, LSTMs use a specialized structure of memory cells with three interacting gates: the input gate, the forget gate, and the output gate. These gates regulate the flow of information into, out of, and within the memory cell.

Input Gate determines how much of the added information should be stored in the memory cell. It is influenced by the current input and the previous output. Forget Gate decides what information from the memory cell should be discarded or kept. It considers the previous output and the current input. Memory Cell stores and maintains the information over extended periods. It is regulated by the input and forget gates. Output Gate decides the next output based on the current input, the previous output, and the information stored in the memory cell.

The strength of LSTMs lies in their ability to retain essential information over extended sequences, making them well-suited for tasks like time series prediction, language modeling, and other applications involving sequential patterns. The intricate interplay of these gates allows LSTMs to selectively store and retrieve information, mitigating the challenges associated with capturing long-range dependencies in sequential data.

**4.2. Risk Factor Calculation**

Our methodical approach to risk factor computation goes beyond the norm to provide a customized investment plan that aligns with each person's financial goals. Now let us examine the detailed procedures that make up our customized risk-based investing approach.

**4.2.1. Customer Input: Critical amount and monthly investments.**

When a client gives us their financial foundation—a lump sum payment named Critical amount and a commitment to monthly investments—our adventure officially begins. We divide this one-time payment (called the critical amount) and monthly investments among several baskets, each of which stands for a different financial goal, to accommodate their individual financial objectives.

**4.2.2. Baskets and Financial Objectives**

We see the customer's financial path from the perspective of various baskets, each of which stands for a different financial goal. These baskets contain valuable information including the anticipated amount, the number of years till maturity, and the family's financial situation. By doing this, we establish the foundation for a detailed comprehension of the client's financial situation.

**4.2.3. Decision Tree Modelling**

In evaluating and judging the decision tree model's performance, we will primarily focus on accuracy metrics tailored to the financial context and the risk factors involved. The accuracy of the model will be assessed by comparing the predicted risk levels (output) with the actual risk levels observed in real-world financial scenarios.

Specifically, the model's performance will be measured using metrics such as precision, recall, and F1 score. Precision is the ratio of true positive predictions to the total number of positive predictions, providing insight into the model's ability to correctly identify high-risk scenarios. Recall, on the other hand, measures the ratio of true positives to the total number of actual high-risk instances, indicating the model's effectiveness in capturing all relevant risky situations.

The F1 score, a harmonic mean of precision and recall, provides a balanced assessment of the model's overall predictive capability. Additionally, metrics like area under the receiver operating characteristic (ROC) curve and the confusion matrix can offer a comprehensive understanding of the model's performance across different risk thresholds.

It's crucial to consider the implications of false positives and false negatives in the financial context. A false positive could lead to an overly cautious risk assessment, potentially causing missed investment opportunities, while a false negative might expose the individual to more risk than they are comfortable with. Therefore, striking a balance between precision and recall is essential in optimizing the decision tree model for effective risk assessment in financial scenarios.

Regularly updating and retraining the model using new financial data will also be a part of the evaluation process to ensure its continued relevance and accuracy in reflecting the dynamic nature of financial markets and individual financial statuses.

**4.2.4. Diversifying the investment money into each basket**

We use a simple approach to use a compound interest formula to determine the principal amount to be invested in each basket. Sorting the baskets according to years to maturity in ascending order starts the diversification process. The amount of principle needed to reach the financial objective for each basket is determined using the estimated return that has been computed. To achieve optimal diversification, we gradually distribute portions of the critical amount among baskets according to their maturity. Until the critical amount is completely used, this process keeps going. The compound interest formula is used to determine the principal number of monthly investments with longer maturity periods. By using an iterative process, the funds are distributed in a way that best suits the risk profile and financial goals of the customer. For monthly investments with longer maturity periods, we employ compound interest formulas. This dynamic approach acknowledges the regular inflow of investments, optimizing the calculation of principal amounts and fostering sustainable growth over time.

We continuously evaluate the viability of the suggested diversification in the spirit of openness. In the event of difficulties, we proactively communicate with clients and suggest modifying the maturity dates or investment amounts. This cooperative discussion makes sure that the investment strategy stays reasonable and in line with the customer's changing financial situation.

Within the group of supervised learning is the machine learning algorithm K-means [44]. Clustering data points according to how close they are to other data points in the feature space works on the similarity principle. When choosing mutual funds, K-means evaluates the past performance of each fund and finds those that most closely conform to our portfolio characteristics.

In conclusion, incorporating K-means into our investment approach is a critical first step in helping our clients maximize their returns. K-means makes sure that the mutual funds chosen by clients meet their overall investment goals in addition to their anticipated returns by utilizing data-driven decision-making and flexibility in response to market conditions. This strategy strengthens our commitment to providing individualized and knowledgeable investment solutions while also optimizing the likelihood of attaining desired results.

**4.3. Classification Model for Stock Financials.**

In the context of our comprehensive stock valuation approach employing diverse models, namely decision tree, random forest, AdaBoost, and XGBoost, a meticulous consideration is accorded to the determination of classification thresholds. While conventional practice often designates a threshold of 0.5 as a binary discriminator, our strategy deviates by establishing a more conservative threshold of 0.6. This intentional adjustment aims to mitigate the incidence of Type I errors, specifically the misclassification of suboptimal stocks as favorable investment opportunities.

Type I errors in this financial context would signify the erroneous classification of stocks as valuable when their underlying financial metrics do not align with prudent investment criteria. By elevating the threshold to 0.6, we deliberately err on the side of caution, prioritizing precision over recall in our predictive models. This strategic choice is grounded in the imperative to minimize the risk associated with false positives, thereby averting the inclusion of financially precarious stocks in the recommended portfolio.

While this decision inevitably introduces the potential for Type II errors, wherein some genuinely promising stocks may be overlooked or excluded, the overarching objective is to safeguard the integrity of investment recommendations by circumspectly curtailing the acceptance of stocks that may harbor undue financial risks. This nuanced calibration of classification thresholds seeks a judicious balance between capturing genuine investment opportunities and fortifying the portfolio against the inclusion of stocks with an elevated likelihood of underperformance or financial instability.

**4.4. Stock Returns Prediction**

This module consists of five sequential steps. In the initial step, the raw data is partitioned into distinct study periods, each further divided into a training segment (utilized for in-sample trading) and a trading segment (employed for out-of-sample predictions). Subsequently, the second step introduces features, while the third step defines the target variables. The fourth step involves setting up CuDNN-LSTM [48]. Finally, in the fifth step, a trading strategy is established for the trading segment.

**4.4.1. Study Periods**

Following the methodology outlined by Krauss et al. [46] and Fischer & Krauss [47], the dataset spanning 29 years from January 1990 to December 2018 is divided into non-overlapping study periods, using a 4-year window and 1-year stride. Each study period consists of a training period of approximately 756 days (3 years) and a trading period of about 252 days (approximately 1 year), resulting in twenty-six study periods with non-overlapping trading segments.

The number of days in a study period is denoted as Tstudy, and ni represents the number of stocks with complete historical data at the end of each study period i. The adjusted closing price cp(s)t and opening price op(s)t of a stock s at time t are defined.

For a prediction day t, the inputs include historical opening prices op(s)t and adjusted closing prices cp(s)t. The task is to predict k stocks with the highest and lowest return.

irt,0: = ,out of all n stocks.

**4.4.2. Feature Generation**

Following the approach of Fischer & Krauss [47], a multi-feature setting is adopted for LSTM. The model is fed with 240 timesteps and three features, trained to predict the direction of the 241st intraday return. Standardization is applied using Robust Scaler, and overlapping sequences of standardized features are generated for each time t.

**4.4.3. Train-Test Split**

For each stock s ϵ S, a matrix with M columns and Tstudy rows is created, where M represents the number of features, and Tstudy is the number of days in a study period (typically 1008 days). This matrix is filled with the respective features, and the top 240 rows are partially filled and subsequently removed since irtm is not defined when t ≤ m. The remaining rows (from t = 241 to Tstudy) are split into two parts: from t = 241 to t = 756 for training and from t = 757 to t = Tstudy for testing. The training data for all stocks in S is concatenated to form the collective training set, while the testing set is created similarly.

**4.4.4. Model Specification for LSTM**

The model consists of twenty-five cells of CuDNN-LSTM, followed by a dropout layer of 0.1 and a dense layer with two output nodes and a SoftMax [49] activation function. The loss function is ordinary least squared [50], and RMSProp [51] is employed as the optimizer with a batch size of 512. Early stopping with a patience of ten epochs, monitoring the validation loss, and a validation split of 0.2 are applied.

The probability P(s)t for each stock s to outperform the median return ir(s)t, zero is forecasted. The trading strategy involves investing heavily in the top k = 10 stocks with the highest P(s)t, all with weightage decided by a regression model with custom loss function we will discuss later.

**4.5. Portfolio Diversification**

The Portfolio Diversification section describes a step-by-step process for selecting and managing investments to reduce risk and potentially increase returns.

**4.5.1. Stock Beta Values and Returns Calculation**

Identification of the beta values for each stock in the portfolio from the current dataset. The beta value measures a stock's volatility compared to the market. This helps in assessing the risk level of each stock.Then, we calculate the historical returns for each stock based on an investment of $100 from section 4.3. This standardizes the return calculations and allows for an easier comparison across different stocks.

**4.5.2. Linear Regression Analysis**

Executing a linear regression model utilizing historical returns data, this framework endeavors to establish a line that prognosticates forthcoming returns predicated on antecedent performance.

Linear regression, a statistical technique, is employed for the modeling and scrutiny of connections between a reliant variable and one or more independent variables. The primary objective is the determination of a linear function that optimally forecasts the values of the dependent variable based on the independent variables. The relational structure is epitomized through a linear function:

*Y* = *β*0​ + *β*1​*X* + *ϵ*

Y is the dependent variable we want to predict.

X is the independent variable used to make predictions.

β0 is the y-intercept of the regression line.

β1 is the slope of the regression line, indicating the change in the dependent variable for each unit change in the independent variable.

ϵ represents the error term, which is the part of y that the linear model cannot explain.

In this model, each stock's return is represented by a variable, here referred to as *Xi*, which stands for the return of the ith stock. The financial beta of each stock is represented as

For the Linear regression model, the relationship between the stock returns per $100 and the predicted returns per $100 is given by:

***R*** =

***R*** =

**4.5.3. Custom Loss Function Application**

**Desired Outcomes:** The key goal in this section is to achieve:

1.

2.

By achieving these, conservative model that prefers underestimating returns over overestimating them to avoid taking on too much risk and ensure that the predictions are realistic and that the model is not overly pessimistic, which could lead to missed investment opportunities.

A loss function is integrated into the regression model to quantify the discrepancies between the model's predictions and actual returns. This function can be tailored to align with specific investment objectives or risk tolerances.

**Squared Error Term :** The first term is the squared difference between the predicted return and the actual return *y*.

This is a term that penalizes larger errors more severely than smaller ones. The goal of minimizing this term is to ensure the model's predictions are as close as possible to the actual returns.

**Regularization Term ​:** The second term ​ acts as a form of regularization, which serves to control the predicted value not being exactly equal to the actual value, as we want a surplus of returns to offset any unexpected losses later. It also keeps our beta as high as possible, which makes sure our model is diversified and always better than the market average. It inversely weighs the prediction error by the product of the error and ***B***, which represents the sum of the products of stock returns per $100 and their respective predicted returns per $100.

**Minimizing Risk:** The goal of the loss function is to minimize the risk by reducing the discrepancy between the predicted () and actual returns, while considering the risk-return profile of the stocks (through B).

**4.5.4. Returns Prediction and Investment Diversification**

The linear regression model, coupled with the loss function, is utilized to forecast the future returns of stocks. This prediction aims to estimate the potential returns from investments in each stock.Based on the regression coefficients—which show the influence of each stock on the predicted returns—we adjust the amount invested in each stock. This step involves allocating more funds to stocks with higher expected returns and lower risk (as per their beta values), and less funds to riskier or less profitable stocks. The diversified investment amounts are determined by how much each stock's predicted return (from the regression analysis) will contribute to the portfolio's overall performance.

**4.5.5. Utilizing Euclidean Distance for Mutual Fund Selection**

In the ever-changing sphere of investments, meeting client-specified return targets can pose challenges, particularly when relying solely on individual stocks. To address this, our investment strategy incorporates a nuanced approach to enhance the attainment of desired expected returns when available stocks fall short of expectations. Our methodology relies on the utilization of the Euclidean distance metric to identify the most suitable mutual funds based on portfolio characteristics.

By employing Euclidean distance, we assess the spatial relationship between our portfolio attributes and those of available mutual funds. This metric serves as a reliable measure of similarity, allowing us to select mutual funds that closely align with the specified return goals. This approach enhances the precision of fund selection, ensuring that the recommended mutual funds effectively meet the targeted return amounts over the investment's maturity period.

The strategic use of Euclidean distance fortifies our investment strategy. It provides a flexible and adaptive framework to optimize the attainment of client-specified return targets. This approach allows us to navigate the complexities of the investment landscape, offering a resilient and tailored portfolio construction that extends beyond individual stocks to encompass a diversified selection of well-aligned mutual funds.

**V. Results**

**5.1 Risk Factor Modeling**

In our exploration of predicting risk factor levels (categorized as 1, 2, or 3) based on family financial status variables—marital status, presence of dependents, income, and age—we employed the Decision Tree and Random Forest models. The respective performance metrics for each model are as follows:

The Decision Tree model exhibited an accuracy of 92.87%, precision of 91.3%, and recall of 93.3%. In contrast, the Random Forest model displayed superior performance, boasting an accuracy of 95.64%, precision of 96.1%, and recall of 95.8%.

The variables used for prediction—marital status, age, presence of dependents, and income—were considered in terms of their importance. In ascending order, the hierarchy of importance is as follows: marital status, age, presence of dependents, and income. This signifies that income holds the most pivotal role, followed by age, presence of dependents, and finally, marital status.

The inclusion of age as a significant variable underscores its importance in predicting risk factor levels. Positioned between the number of dependents and income in terms of importance, age introduces a nuanced dimension to the predictive models.

The preference for the Random Forest model remains, given its ensemble nature, which combines multiple decision trees to provide a more robust and accurate predictive framework. This is especially valuable in scenarios where individual decision trees may be susceptible to overfitting. In table 1 and Figure 15, you can find the model results and variable importance, respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| Decision Tree | 0.9287 | 0.913 | 0.933 |
| Random Forest | 0.9564 | 0.961 | 0.958 |

## Table 1 – Risk Factor Prediction Results.

A graph with text on it

Description automatically generated

## Figure 15 – Variable Importance Plot for Random Forest Risk Factor Prediction.

**5.2 Interest Rate Prediction**

In the examination of models predicting interest rates based on risk factor, amount needed at maturity, and years until maturity, a suite of regression models was considered, encompassing Linear Regression, Lasso Regression, Ridge Regression, and XGBoost. The evaluation was grounded in the Root Mean Squared Error (RMSE) metric, serving as a quantitative measure of the average deviation between predicted and observed interest rates.

Linear Regression, as the foundational model, provided a baseline prediction with an RMSE of 0.58. While offering a rudimentary understanding, this model demonstrated limitations in capturing the intricate relationships embedded in the dataset. Introducing regularization techniques, Lasso Regression, designed to discourage unnecessary predictors, exhibited an improved performance with an RMSE of 0.45. However, challenges persisted in fully elucidating the complex relationships within the data. Ridge Regression, leveraging regularization to address multicollinearity issues, further enhanced predictive capability, yielding an RMSE of 0.39. Despite improvement over both Linear and Lasso Regression, Ridge displayed constraints in fully capturing the nuanced nature of the underlying dataset.

In stark contrast, the XGBoost model emerged as the most effective, achieving a remarkably low RMSE of 0.23. Harnessing the power of ensemble learning through gradient boosting, XGBoost displayed superior performance in discerning intricate non-linear patterns within the dataset. This notable outcome positions XGBoost as the optimal model for predicting interest rates in this context.

The observed performance differences underscore the inherent complexity of the dataset and the varying capacities of regression models to address such intricacies. While Linear Regression provides a fundamental insight, the introduction of regularization techniques enhances predictive accuracy. XGBoost, with its ensemble nature, proves instrumental in surpassing the predictive capabilities of all other models, signifying its efficacy in capturing the nuanced relationships inherent in predicting interest rates based on risk factor, amount needed at maturity, and years until maturity. The individual model results are summarized in Table 2.

A table with numbers and letters

Description automatically generated

## Table 2 – Interest Rate Prediction Models.

**5.3 Stock Classification**

In the examination of stock classification models based on financial indicators, including CART (Classification and Regression Trees), Logistic Regression, kNN (k-Nearest Neighbors), and XGBoost, the models present varying performances in terms of accuracy and false positive rates.

CART, with an accuracy of 76.42%, demonstrates a higher overall accuracy but is associated with a notable false positive rate of 21.59%. This implies instances where the model incorrectly identifies stocks as positive, raising concerns about its precision in classification.

Logistic Regression, with an accuracy of 70.85%, exhibits a higher false positive rate of 35.50%. The increased rate of false positives is a significant consideration, emphasizing potential challenges in accurately distinguishing positive instances in stock classification.

kNN, achieving an accuracy of 75.30%, is characterized by a false positive rate of 37.51%. Like Logistic Regression, this model faces challenges in minimizing false positives, indicating potential limitations in the reliability of its positive classifications.

In contrast, XGBoost stands out as the top performer with an accuracy of 95.55% and an impressively low false positive rate of 2.80%. The model demonstrates a robust ability to accurately classify stocks while notably minimizing the risk of false positives, making it a promising choice for stock classification tasks.

In conclusion, the trade-off between accuracy and false positive rates is crucial in stock classification models. XGBoost, with its exceptional accuracy and minimal false positive rate, emerges as a strong candidate, highlighting a balance between precision and reliability. This underscores the potential efficacy of XGBoost in stock classification, particularly when prioritizing the reduction of false positive instances. The results can be seen below in Table 3.

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **False Positive Rate** |
| CART | 0.7642276 | 0.2159412 |
| Logistic Regression | 0.7085020 | 0.3550362 |
| kNN | 0.7530364 | 0.3750780 |
| XGBoost | 0.9555 | 0.028 |

## Table 3 – Stock Classification based on company Financials – Model Results.

**5.4 Stock Price Forecasting**

In the context of utilizing Long Short-Term Memory (LSTM) networks for predicting the Compound Annual Growth Rate (CAGR) for a portfolio consisting of three thousand stocks over the next five years, the evaluation included an analysis of memory utilization and garbage collection metrics. The computational workload reached its peak, utilizing 10996121 Mb for Ncells and 87042338 Mb for Vcells.

Simultaneously, the predictive accuracy of the LSTM model was assessed by examining the loss at different half-yearly time steps across the three thousand stocks. Notably, a substantial increase in loss was observed at the fourth time-step, corresponding to the 2-year mark. This significant shift in loss patterns prompted a strategic decision to evaluate the model's performance at intervals every 2 years. This periodic assessment aims to ensure continued alignment with projections and to identify and address potential deviations from expected performance. By implementing this evaluation strategy, the objective is to maintain the model's effectiveness in predicting CAGR for the portfolio of three thousand stocks over the entire forecasting period.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **used** | **(Mb)** | **gc trigger** | **(Mb)** | **max used** | **(Mb)** |
| **Ncells** | 7585337 | 405.2 | 10996121 | 587.3 | 10996121 | 587.3 |
| **Vcells** | 52687163 | 402.0 | 87042338 | 664.1 | 87042274 | 664.1 |
|  |  |  |  |  |  |  |

## Table 4 – LSTM Details.

|  |  |
| --- | --- |
| **Time-step** | **Loss** |
| 1 | 0.05 |
| 2 | 0.21 |
| 3 | 0.33 |
| 4 | 1.76 |
| 5 | 8.23 |
| 6 | 16.89 |
| 7 | 27.36 |
| 8 | 57.01 |

## Table 2 – Loss by Timestep – averaged over three thousand stocks.

**5.5 Linear Regression for Portfolio Diversification**

In the refinement of our portfolio optimization process, we introduced a custom loss function as outlined in the methodology. This loss function was specifically designed to align predicted returns with expected returns obtained from the initial step utilizing the Random Forest model for 3000 stocks. The optimization procedure involved differentiating this loss function over the coefficients, representing the proportion of principal amount allocated to each stock, and subsequently equating the derivative to zero to identify the minimal value.

The uniqueness of this approach lies in its self-contained nature, eliminating the need for external benchmarks or comparison sources. By focusing on the internal consistency of the loss function, our optimization method ensures that the expected returns are effectively matched, facilitating the attainment of an optimal stock diversification strategy. It is important to note that the custom loss function enforces a constraint where the sum of coefficients for all stocks adds up to 1, inherently addressing the proportionality of the principal amount allocated across the portfolio.

While specific numerical results are omitted here, the assurance of expected return alignment and the achievement of an optimal stock diversification through coefficient optimization underscore the efficacy of our approach in tailoring investment strategies based on the predicted returns and ensuring the coherence of the portfolio's composition.

**VI. Conclusion**

In conclusion, this research endeavors to revolutionize investment management by harnessing the power of predictive algorithms and advanced computational models. Focused on addressing the challenges posed by the evolving financial landscape and the variability of individual financial needs, our approach aims to provide a pathway to financial success through data-driven investments.

The project leverages a combination of strategic models, such as Random Forest for return predictions and classification inspired by Warren Buffett's value investing strategy, to construct tailored investment portfolios for users. These portfolios are designed to align with individual financial goals, risk preferences, and time horizons. By utilizing sophisticated techniques, including Long Short-Term Memory (LSTM) networks and customized regression models, we seek to bridge the gap between traditional financial strategies and the transformative potential of machine learning.

Central to our methodology is the recognition of the nuanced relationship between investor profiles and market dynamics. The integration of decision trees for risk tolerance, neural networks for efficient portfolio diversification, and regression models for predictive market behavior underscores our commitment to providing comprehensive and personalized investment recommendations.

This research represents a significant advancement in the convergence of finance and technology, aiming to empower individuals on their quest for financial freedom. By offering a platform that transcends traditional barriers and provides access to tailored investment strategies, we aspire to democratize financial decision-making. It is imperative to note that while our research is grounded in predictive modeling algorithms, it should not be construed as financial advice without context or verification by a professional.

In summary, through the synthesis of data science and financial principles, our research seeks to encourage individuals to embark on their journey toward financial success, fostering a future where data-driven investments play a pivotal role in achieving diverse financial aspirations.

**VII. Future Work**

In charting the course for future endeavors, this research acknowledges several promising avenues for refinement and expansion. A pivotal focal point for prospective exploration resides in the sphere of Long Short-Term Memory (LSTM) models, constituting a foundational element within our comprehensive methodology.

To begin, the refinement of LSTM models emerges as a priority. This entails a more exhaustive exploration of network architecture variations, hyperparameter fine-tuning, and the incorporation of additional pertinent features to enhance the precision and robustness of our forecasting framework.

Another dimension of future work involves the integration of external factors into our predictive models. The inclusion of macroeconomic indicators, geopolitical events, and global market trends stands out as a potential enhancement to offer a more comprehensive understanding of market dynamics and potentially elevate the accuracy of stock predictions.

Additionally, the realm of risk management strategies presents an avenue for development. The refinement of decision tree models to better assess and adapt to changes in market volatility could fortify our risk assessment mechanisms, ensuring their adaptability to the evolving landscape of financial conditions.

Exploring alternative machine learning models constitutes another potential trajectory. Beyond our existing repertoire, investigating ensemble methods or deep learning architectures may reveal novel insights and avenues for augmenting the overall efficacy of predictive algorithms.

The development of a user-friendly interface for seamless interaction with predictive models and investment recommendations represents a crucial aspect of future work. Such an interface would empower users to input and update their financial information, track portfolio performance, and receive real-time insights in a more intuitive manner.

Furthermore, the integration of real-time data feeds emerges as a key consideration. Incorporating mechanisms to continuously update datasets and swiftly adapt to changing market conditions can enhance the timeliness and relevance of investment recommendations.

Validation against external benchmarks and established financial metrics constitutes an essential facet of ongoing research efforts. This rigorous validation process, extending beyond the metrics discussed in this research, ensures a more thorough assessment of the model's predictive capabilities and alignment with proven financial methodologies.

Notably, the module centered on LSTM-based stock predictions presents a particularly fertile ground for future exploration. The continuous refinement of this domain holds the potential to unlock new insights, patterns, and innovations in predictive analytics, thereby ensuring the sustained relevance and efficacy of our algorithms in the dynamic landscape of stock market forecasting.

**References**

[1] Sirignano, Justin & Cont, Rama. (2019). Universal features of price formation in financial markets: perspectives from deep learning. Quantitative Finance. 19. 1-11. <https://doi.org/10.1080/14697688.2019.1622295>

[2] Eunsuk Chong, Chulwoo Han, Frank C. Park, Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies, Expert Systems with Applications, Volume 83, 2017, Pages 187-205, ISSN 0957-4174. <https://doi.org/10.1016/j.eswa.2017.04.030>

[3] Thomas Fischer, Christopher Krauss, Deep learning with long short-term memory networks for financial market predictions, European Journal of Operational Research, Volume 270, Issue 2, 2018, Pages 654-669, ISSN 0377-2217. <https://doi.org/10.1016/j.ejor.2017.11.054>

[4]  [Bao, Wei & Yue, Jun & Rao, Yulei. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE. 12. https://doi.org/10.1371/journal.pone.0180944](file:///C:\Users\sc30a\Python%20Projects\Investment_Management\Paper\%20Bao,%20Wei%20&%20Yue,%20Jun%20&%20Rao,%20Yulei.%20(2017).%20A%20deep%20learning%20framework%20for%20financial%20time%20series%20using%20stacked%20autoencoders%20and%20long-short%20term%20memory.%20PLoS%20ONE.%2012.%20https:\doi.org\10.1371\journal.pone.0180944%20%20)

[5] Hum Nath Bhandari, Binod Rimal, Nawa Raj Pokhrel, Ramchandra Rimal, Keshab R. Dahal, Rajendra K.C. Khatri, Predicting stock market index using LSTM, Machine Learning with Applications, Volume 9. <https://doi.org/10.1016/j.mlwa.2022.100320>

[6] [Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. https://doi.org/10.1162/neco.1997.9.8.1735.](file:///C:\Users\sc30a\Python%20Projects\Investment_Management\Paper\Hochreiter,%20Sepp%20&%20Schmidhuber,%20Jürgen.%20(1997).%20Long%20Short-term%20Memory.%20Neural%20computation.%209.%201735-80.%20https:\doi.org\10.1162\neco.1997.9.8.1735.%20%20)

[7] Bailey, David H. and López de Prado, Marcos and López de Prado, Marcos, The Sharpe Ratio Efficient Frontier (April 1, 2012). Journal of Risk, Vol. 15, No. 2, Winter 2012/13, [http://dx.doi.org/10.2139/ssrn.1821643](https://dx.doi.org/10.2139/ssrn.1821643)

[8] Bank, D., Koenigstein, N., Giryes, R. (2023). Autoencoders. In: Rokach, L., Maimon, O., Shmueli, E. (eds) Machine Learning for Data Science Handbook. Springer, Cham. <https://doi.org/10.1007/978-3-031-24628-9_16>

[9] 6. D. Karmiani, R. Kazi, A. Nambisan, A. Shah and V. Kamble, "Comparison of Predictive Algorithms: Backpropagation, SVM, LSTM and Kalman Filter for Stock Market," 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 2019, pp. 228-234. <https://doi.org/10.1109/AICAI.2019.8701258>

[10] Bhansali, R. J. “Linear Prediction by Autoregressive Model Fitting in the Time Domain.” The Annals of Statistics 6, no. 1 (1978): 224–31. <http://www.jstor.org/stable/2958705>

[11] Christian Breitung, Automated stock picking using random forests, Journal of Empirical Finance, Volume 72,2023, Pages 532-556, ISSN 0927-5398

[12] Alsulmi, M. (2022). From Ranking Search Results to Managing Investment Portfolios: Exploring Rank-Based Approaches for Portfolio Stock Selection. Electronics, 11(23), 4019.

[13] S. Saha, J. Gao, and R. Gerlach, "Stock Ranking Prediction Using List-Wise Approach and Node Embedding Technique," in IEEE Access, vol. 9, pp. 88981-88996, 2021, doi: 10.1109/ACCESS.2021.3090834.

[14] Qiang Song, Anqi Liu, Steve Y. Yang, Stock portfolio selection using learning-to-rank algorithms with news sentiment, Neurocomputing, Volume 264,2017, Pages 20-28, ISSN 0925-2312

[15] Robert J. Yan and Charles X. Ling. 2007. Machine learning for stock selection. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '07). Association for Computing Machinery, New York, NY, USA, 1038–1042. <https://doi.org/10.1145/1281192.128130>

[16] C. -F. Huang, T. -N. Hsieh, B. R. Chang and C. -H. Chang, "A comparative study of stock scoring using regression and genetic-based linear models," 2011 IEEE International Conference on Granular Computing, Kaohsiung, Taiwan, 2011, pp. 268-273, <https://doi.org/10.1109/GRC.2011.6122606>.

[17] Hota, J., Chakravarty, S., Paikaray, D., & Bhoyar, H. (2022). Stock Market Prediction Using Machine Learning Techniques. ACI@ISIC.

[18] Matthias X. Hanauer, Tobias Kalsbach, Machine learning and the cross-section of emerging market stock returns, Emerging Markets Review, Volume 55,2023,101022, ISSN1566-0141.

[19] JOUR, Saha, Suman, Gao, Junbin, Gerlach, Richard, 2022, A survey of the application of graph-based approaches in stock market analysis and prediction, International Journal of Data Science and Analytics, Volume - 14.

[20] Tin Kam Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995, pp. 278-282 vol.1, https://doi.org/ 10.1109/ICDAR.1995.598994.

[21] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank from pairwise approach to listwise approach. In Proceedings of the 24th international conference on Machine learning (ICML '07). Association for Computing Machinery, New York, NY, USA, 129–136. <https://doi.org/10.1145/1273496.1273513>

[22] Hosein, N.; Kheyroddin, A.M.S. Risk Assessment in Bridge Construction Projects in Iran Using Monte Carlo Simulation Technique. Pract. Period. Struct. Des. Constr. 2019, 24, 04019026. <https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29SC.1943-5576.0000450>

[23] Dheskali, E.; Koutinas, A.A.; Kookos, I.K. Risk assessment modeling of bio-based chemicals economics based on Monte-Carlo simulations. Chem. Eng. Res. Des. 2022, 163, 273–280. <https://doi.org/10.1016/j.cherd.2020.09.011>

[24] Platon, V.; Constantinescu, A. Monte Carlo Method in Risk Analysis for Investment Projects. Procedia Econ. Finance. 2014, 15, 393–400. <https://doi.org/10.1016/S2212-5671(14)00463-8>

[25] Albogamy, A.; Dawood, N. Development of a client-based risk management methodology for the early design stage of construction processes. Eng. Constr. Archit. Manag. 2015, 22, 493–515. <https://doi.org/10.1108/ECAM-07-2014-0096>

[26] Arnold, U.; Yildiz, Ö. Economic risk analysis of decentralized renewable energy infrastructures – A Monte Carlo Simulation approach. Renew. Energy 2015, 77, 227–239. <https://doi.org/10.1016/j.renene.2014.11.059>

[27] Kumar, L.; Jindal, A.; Velaga, N.R. Financial risk assessment and modelling of PPP based Indian highway infrastructure projects. Transp. Policy 2018, 62, 2–11. <https://doi.org/10.1016/j.tranpol.2017.03.010>.

[28] Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 855–864. <https://doi.org/10.1145/2939672.2939754>

[29] Mirjalili, S. (2019). Genetic Algorithm. In: Evolutionary Algorithms and Neural Networks. Studies in Computational Intelligence, vol 780. Springer, Cham. <https://doi.org/10.1007/978-3-319-93025-1_4>

[30] Zaimovic, Azra, Adna Omanovic, and Almira Arnaut-Berilo. 2021. "How Many Stocks Are Sufficient for Equity Portfolio Diversification? A Review of the Literature" Journal of Risk and Financial Management 14, no. 11: 551. <https://doi.org/10.3390/jrfm14110551>

[31] Ahmadi-Javid, A., Fateminia, S. H., & Gemünden, H. G. (2020). A Method for Risk Response Planning in Project Portfolio Management. Project Management Journal, 51(1), 77-95. <https://doi.org/10.1177/8756972819866577>

[32]Myles, Anthony J., Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. "An introduction to decision tree modeling." *Journal of Chemometrics: A Journal of the Chemometrics Society* 18, no. 6 (2004): 275-285.

[33] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*, 199-222.

[34] Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31-44.

[35] Heaton, J. B., N. G. Polson, and J. H. Witte. 2017. Deep learning for ﬁnance: Deep portfolios. Applied Stochastic Models in Business andIndustry 33: 3–12 <https://onlinelibrary.wiley.com/doi/10.1002/asmb.2209>

[36] Durairaj, M., and B. H. Krishna Mohan. 2019. A review of two decades of deep learning hybrids for ﬁnancial time series prediction. International Journal on Emerging Technologies 10: 324–331.

[37] Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71-72.

[38] Pitman, J. (1999). Coalescent random forests. *Journal of Combinatorial Theory, Series A*, *85*(2), 165-193.

[39] Schapire, R. E. (1999, July). A brief introduction to boosting. In *Ijcai* (Vol. 99, No. 999, pp. 1401-1406).

[40] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

[41] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[42] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[43] Medsker, L., & Jain, L. C. (Eds.). (1999). *Recurrent neural networks: design and applications*. CRC press.

[44] Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, 80716-80727.

[45] Velimirović, D., Velimirović, M., & Stanković, R. (2011). Role and importance of key performance indicators measurement. *Serbian Journal of Management*, *6*(1), 63-72.

[46] Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689-702.

[47] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European journal of operational research*, *270*(2), 654-669.

[48] Braun, S. (2018). LSTM benchmarks for deep learning frameworks. *arXiv preprint arXiv:1806.01818*.

[49] Gold, S., & Rangarajan, A. (1996). Softmax to softassign: Neural network algorithms for combinatorial optimization. *Journal of Artificial Neural Networks*, *2*(4), 381-399.

[50] Van Huffel, S., & Zha, H. (1993). 10 The total least squares problem.

[51] Liu, J., & Xu, Y. (1991). Network optimization.

**Glossary**

**Mutual Funds**: Mutual funds are a collection of numerous investors' money that they can use to buy stocks, bonds, or other assets. Professional investment companies oversee managing these funds.

**ETF**: Mutual funds and stocks combine to create exchange-traded Funds (ETFs). Like mutual funds, they facilitate easy investment into a wide range of assets, but unlike mutual funds they trade on stock exchanges like individual stocks.

**Alpha**: Alpha is a term used in the financial industry to evaluate how well a mutual fund or investment portfolio performs in relation to a benchmark index.

Where alpha = Actual return of the fund – Expected return of the fund

**Beta**: A mutual fund or stock's beta value indicates how sensitive it is to changes in the market, especially when it comes to shifts in a benchmark index like the S&P 500. It shows the expected range of return fluctuations of the fund with respect to the market.

Where Beta= Covariance of fund returns with market returns/variance of market returns

**Sharpe Ratio**: It is a tool for calculating an investment or portfolio's risk-adjusted return is the Sharpe ratio. Assessing whether the additional return on an investment justifies the increased risk is one of its functions.

Where Sharpe ratio= (return of investment- risk free-rate)/ Standard Deviation of Investment

**Sortino Ratio**: The Sortino Ratio represents the risk-adjusted return on an investment. It is like Sharpe ratio but rather than concentrating on the investment's overall volatility, it particularly addresses the downside risk—that is, the risk of losing money.

Where Sharpe ratio= (return of investment- risk free-rate)/ Downside Deviation of Investment

**CAGR**: Compound Annual Growth Rate, or CAGR for brief is a way to calculate the growth or decline of an investment over the course of years while considering the compounding effect. In short, the CAGR indicates the average annual percentage return on your investment after reinvested gains or losses from prior years.

Where CAGR= [(Ending Value/Beginning Value) ^ (1-no. of years)]-1

**Expense Ratio**: The cost of managing and running a mutual fund or exchange-traded fund (ETF) is the expense ratio. The standard way to express it is as a percentage of the total assets of the fund.

Where expense ratio= (total annual expenses/average total assets)

**Large cap Funds**: The main investments made by these funds are in large, established companies. These companies are usually the largest and most well-known on the stock market.

They are usually less risky and more stable because they have already grown to a respectable size. Buying stock in the "big players" in the industry is comparable to buying large-cap funds.

**Medium cap funds**: Companies that are larger than small caps but smaller than large caps are the focus of medium-cap funds. They are frequently thought to possess a healthy balance between risk and growth potential. Although they might not be as stable as large caps, companies in this category are typically growing and expanding.

**Small cap funds**: Small-cap funds make investments in recently founded, smaller businesses that are still in the early phases of development. These businesses have room to grow, but because they might not be as established or financially secure as larger businesses, they also carry a higher risk. Small-cap fund investing is like investing in emerging companies.

**Fund Family**: A financial institution or investment management firm that provides a range of mutual funds is called a fund family. These funds frequently have a similar investment philosophy or strategy and are managed by the same organization.

**PE ratio**: The Price-to-Earnings (PE) ratio is a financial ratio used by investors to determine how much a company's stock is worth.

PE ratio=Market price per share/ Earnings per share

**S&P 500**: The S&P 500, also known as the Standard & Poor's 500, is an index of the stock market that evaluates the performance of five hundred of the biggest and most well-known publicly traded companies in the United States.

**NASDAQ**: The American stock exchange NASDAQ is well-known for its electronic trading system and its emphasis on technology and businesses with room to grow. It is among the biggest and most well-known stock exchanges in the world as well as the United States. The National Association of Securities Dealers Automated Quotations is what NASDAQ is.