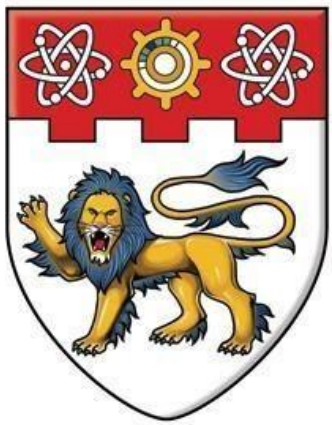


NANYANG TECHNOLOGICAL UNIVERSITY

BC2406 Analytics I: Visual & Predictive Techniques

Group Project Report

Seminar Group No. 5

Team No. 6

Group Members

CHUA WEI JIAN (U1810087C)

NG YAN MING (U1910470H)

TEY CHIN YI (U1920368L)

TING NAI XIANG, MATTHEW (U1922331J)

CONTENT PAGE

Executive Summary

The report aims to tackle the issue of low fund returns resulting from asset managers' bias and imperfect knowledge. This is a pressing problem faced by asset management companies like White Rock and has led to the loss of customers. Ideally, there should be no bias or imperfect knowledge and asset managers are able to recommend the optimal portfolio to their clients. With the aid of analytics, we aim to provide an optimal composition for the portfolio and the process taken to reach it, that is easy to understand. Our choice of models is based on its high explainability power to be able to convince the manager to alter their current compositions.

The dataset used for our analysis provides insight into the performance of funds over 9 years. The use of text-mining to get the general market sentiment for the year and splitting the data to good and bad year, before using the linear regression and Classification and Regression Tree (CART) model, would be important in our prediction of the rate of returns. The initial split using text-mining is vital as the industry sectors would perform differently in good or bad years. Different models will be created depending on the risk-appetite of their clients to allow managers to recommend the portfolio most suitable based on the client's profile.

Linear regression model would then be utilised to provide an estimate of the sectors that managers should invest in and the corresponding weights of each sector. Managers might be tempted to just have one factor or industry in their portfolio, however this is not optimal as the risk of doing that is very high. Hence, we would have to incorporate the CART model.

CART model provides the optimal tree path that the manager should follow. This optimal path coupled with the highest coefficient from the linear regression model would serve as a guide to the asset manager to decide what proportion of industries should his portfolio contain.

With a loss of clients due to suboptimal returns due to poor portfolio composition, it is of paramount importance that White Rock is able solve the issue of poor portfolio composition.

1. Overview of Business Problem

1.1 Problem Statement

Asset Manager's Bias & Imperfect Knowledge

Behavioural bias tends to influence asset managers' decision making process. Stability bias, one type of such bias, occurs when these managers make investment decisions based mostly on their past experiences and knowledge (Hoffman, Huber, & Smith, 2017). There will be instances where managers overlook products that could have potentially yielded higher returns and overweighting on products as a result of their preference or biases. Furthermore, managers lack the time to look through the numerous investment products to form their portfolio, causing imperfect knowledge.

Implications

Familiarity bias occurs when asset managers have a preference for a familiar investment even though there are other viable alternatives for portfolio diversification (Kase, 2018). Investing in an asset that they have owned before provides them with a false sense of security. As such, familiarity bias affects portfolio construction and can lead to suboptimal choices.

Furthermore, asset managers have a tendency to make decisions based on their confidence in self-based knowledge (Kenton, 2020). Self-attribution bias causes them to attribute the success of the fund's performance to themselves making the right choices, thereby boosting their confidence even though they are holding a suboptimal portfolio. However, when the funds are underperforming, they would simply attribute that to external factors and are unwilling to make changes.

Given the sheer volume of financial data that is being generated daily (Nath, 2019), and that asset managers only focus on their biased selection, it is impractical for managers to be aware of all the trends to be able to take advantage of them. This imperfect knowledge, coupled with their confidence in their prior knowledge, will ultimately result in suboptimal portfolio construction.

Asset management companies like White Rock thus face issues with customer retention since the decreasing returns does not justify the fees charged compared to Passive Index Funds, which is gaining more popularity, as they require less fees and provide reasonable outputs. (PwC, 2019)

2.0 Overview of Solution

Tackling manager's bias and imperfect knowledge

By utilising a dataset of fund's performance over a period of 9 years, managers will have a clearer view of the market situation, other fund's performance and composition. This reduces the issue of imperfect knowledge as the dataset contains a list of all major listed funds. Managers will no longer be limited to their personal expertise and knowledge, overcoming the bias identified earlier.

Choosing an area of focus

According to Kiplinger, 30% of the top mutual funds to buy are from the category "Large-Growth" (Waggoner, 2019). Moreover, 5.2% of US mutual funds (Yahoo Finance) belong to the "Large-growth" category.

Since White Rock clientele mainly includes those seeking high returns as they would be paying a management fee, most investment products that White Rock would recommend to their clients would be those that offer higher potential capital appreciation and the associated above-average risk, which fits the description of Large-Growth funds. Hence, the group has decided to analyse Large-Growth Funds as this will appeal to White Rock's clients.

Our research

Our group will be studying the effect of portfolio composition (asset classes) and industry (sector composition) on the predicted rate of return of the fund. Asset classes and industry sectors are selected as our independent variables because managers have control over these two factors and they would have a direct effect on returns.

Current industrial approach

Asset management companies employ a combination of machine learning and artificial intelligence to glean critical insights from data. Natural language processing is used to process public filing and sentiment analysis that would affect the stock. Signals such as business risk, credit ratings, debt transaction and financial filings captured across a variety of sources are being processed to identify new trade patterns. (Kerrigan, Williams, Smith, Petitto, & Nolting, 2020)

However, asset managers are still not receptive to large scale analytics as they are skeptical of new methods involving technology and would rather rely on their knowledge and experience.

Whereas our solution is to incorporate elements of experts' trading activities and opinions so as to improve the credibility of the solution, while allowing space for asset managers' own opinions to be involved in decision making.

Objective:

The prediction model will allow asset managers to determine the return rates of their fund selection so as to aid them in optimising their rate of returns.

This would rule out the manager's bias and imperfect knowledge as our modelling is based on historical data that combines numerous experts' opinions as well as other professionals' trading activities. By leveraging on data of how other professionals are trading, we will be able to identify areas of potential. These key sectors will be highlighted to the attention of the asset manager.

3. Dataset

3.1 Dataset information

https://www.kaggle.com/stefanoleone992/mutual-funds-and-etfs

The data set we used comes from past and existing United States of America (USA) mutual funds available in the market, scrapped from Yahoo! Finance. There are a total of 25,265 mutual funds in this dataset, which provides the following information:

- Fund Name (Symbol and Full Name)
- Category
- Fund Family
- Net Assets
- Returns (Year-to-date (ytd), fund yield)
- Morningstar Rating
- Inception Date
- Investment Type and Size
- Currency
- Net Annual Expense Ratio (Fund & Category)

- Portfolio Weightage (Cash, Stocks, Bonds, Preferred Stocks, Convertible Bonds and Others)
- Price Ratios (Compared to Earnings, Book, Sales, Cashflow)
- Median Market Cap
- Industry Weightage (Basic Materials, Consumer Cyclical, Financial Services, Real Estate, Consumer Defensive, Healthcare, Utilities, Communications, Energy, Industrials and Technology)
- Bond Maturity and Duration
- Ratings of bonds from USA Government
- Bond Ratings (AAA, AA, A, BBB, BB, B, below B and others)
- Morningstar Returns & Risk Rating
- Fund and Category Returns by time (ytd, 1 month, 3 months, 1 year, 3 years, 5 years, 10 years)
- Fund and Category Returns by year (2010 to 2018)
- Number of years up and down
- Fund & Category Alpha & Beta (3 years, 5 years, 10 years)
- Fund & Category Mean Annual Returns (3 years, 5 years, 10 years)
- Fund & Category R-square (3 years, 5 years, 10 years)
- Fund & Category Standard Deviation (3 years, 5 years, 10 years)
- Fund & Category Sharpe Ratio (3 years, 5 years, 10 years)
- Fund & Category Treynor Ratio (3 years, 5 years, 10 years)

3.2 Data Cleaning

Cleaning for Data Exploration

For each dataset, categories other than “Large Growth” will be removed. Also, the data had been cleaned such that the total industry weightage is around 100%. There are 1331 rows after cleaning. After that, each fund needs the total number of sectors that it has invested in and also finds the sector that the fund invested the most by creating a separate dataset. After creating, the number of sectors and the largest proportion sector is added into the original data. There are also some funds that do not have fund returns of 1, 3, 5 and 10 years. Therefore, it is required to remove them for further analysis.

Cleaning for Model Building

For each dataset, categories other than “Large Growth” will be removed. The data is then split into bull or bear markets, which refers to ‘good’ or ‘bad’ market periods. Next, returns, alpha and beta are averaged out from 3, 5, 10 years, for each of the markets, and the NA values from the average are removed. Any outliers of the average alpha are removed to remove any extreme performance. Lastly, the data is further split by beta either less than equal to 1 or above 1.

Columns Removed

As we do not need every single column for the analysis, we will eliminate some of the columns we will not use. The columns removed are as follows:

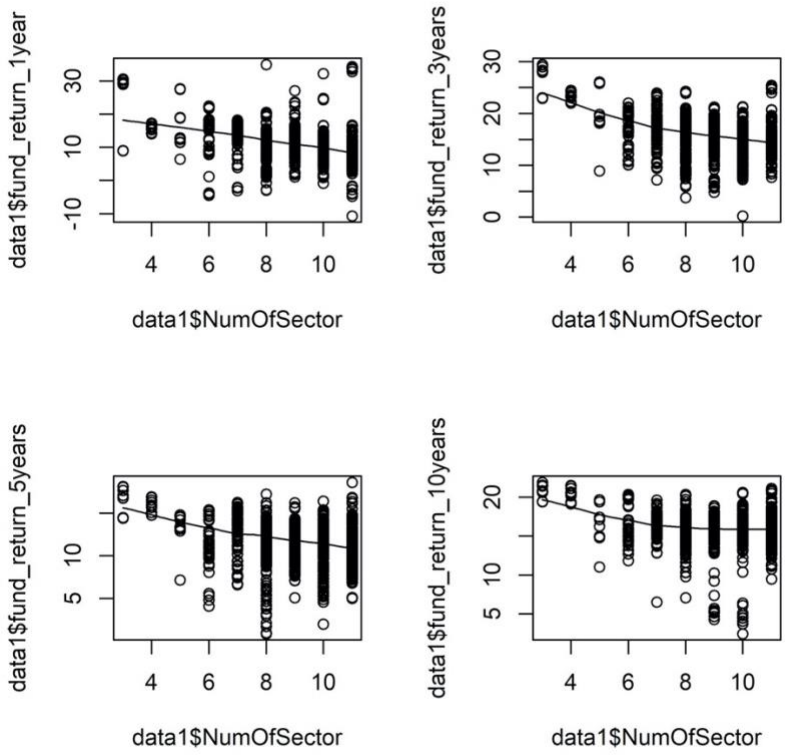
- ***ytd_return*** and ***fund_return_ytd*** are duplicate columns, so we will remove ***ytd_return.***
- Since all the funds are in USD, we do not need the ***currency*** column.
- As we are mainly focusing on stocks, all bond related columns (Appendix A) are removed as they are not required to evaluate portfolios with the majority of stocks.
- We were only comparing based from 2010 to 2018, ***years_up*** and ***years_down*** column is removed as it includes the years before 2010 and we do not have data before 2010.
- Any category related columns (Appendix A) are removed because we are only focusing on 1 type of category, “Large Growth”, which means that all the category returns are constant, making the category columns irrelevant.
- Any MorningStar ratings (Appendix A) are also removed because we are going to beta as our measurement of risk and MorningStar ratings are limited to 1 to 5 stars, which does give a larger depth of risk assessment.
- Any price ratios (Appendix A) are removed as they are dependent from the rate of returns and hence, will not be useful.
- ***net_annual_expense_ratio_fund*** is not related to rate of returns, so we remove it.
- As we are only focusing on Large Growth funds, it makes ***size*** and ***investment*** irrelevant as all size will be large and investment will be growth after categorizing them.

- ***fund_family*** is removed as we are comparing all funds equally and which fund family it comes from does not influence the results.
- ***inception_date*** is removed as we only need 1 year of fund returns for comparison and if there are no returns for 1 year, means the fund is new.
- ***fund_yield*** is removed as we are trying to find the rate of return and the yield only provides part of the return and not the whole picture.
- ***net_assets*** is removed as we are trying the rate of returns in percentage and hence net assets will not influence the percentage change of returns.
- ***median_market_cap*** is removed as it does not affect the rate of returns.

3.3 Data Exploration

Number of sectors invested in against fund return

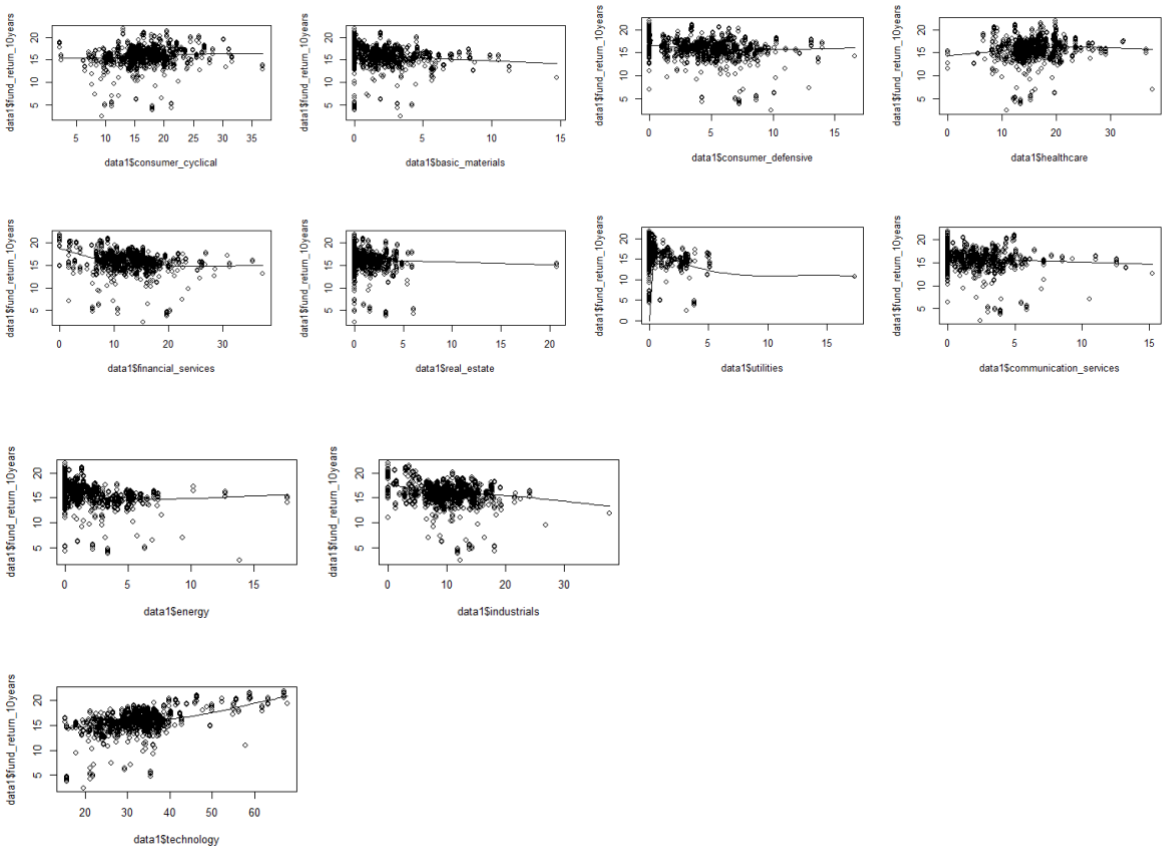
We see a downward trend for all 4 time periods (1,3,5,10 years) as the number of sectors invested increases (Fig 1). This confirms our hypothesis that as we diversify more, the returns would decrease (less profitable).



(Figure 1: Rate of returns compared to number of sectors)

Individual sector percentages against the average return of fund

In general, all sectors show a gentle positive/negative curve except for technology that shows an increase as the percentage of technology in the fund increases (Fig 2).



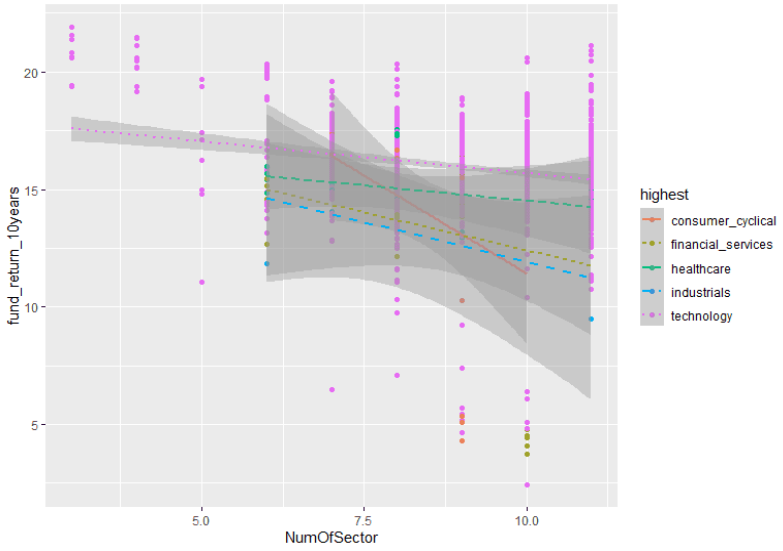
(Figure 2: Rate of Return compared to industry)

Number of sectors invested in against average return by dominant sector

This graph shows the dominant industry for each fund (i.e the sector that makes up the highest percentage of a fund) (Fig 3).

We are able to see that technology dominated funds show the lowest decrease as the number of sectors increases. However, we can conclude that no matter which sector is dominant, the fund tends to earn less as more sectors are invested in.

One possible reason is that asset managers are not employing the optimal diversification strategies. This is supported by the above graphs which shows that the percentage of sector investment does affect the average returns.



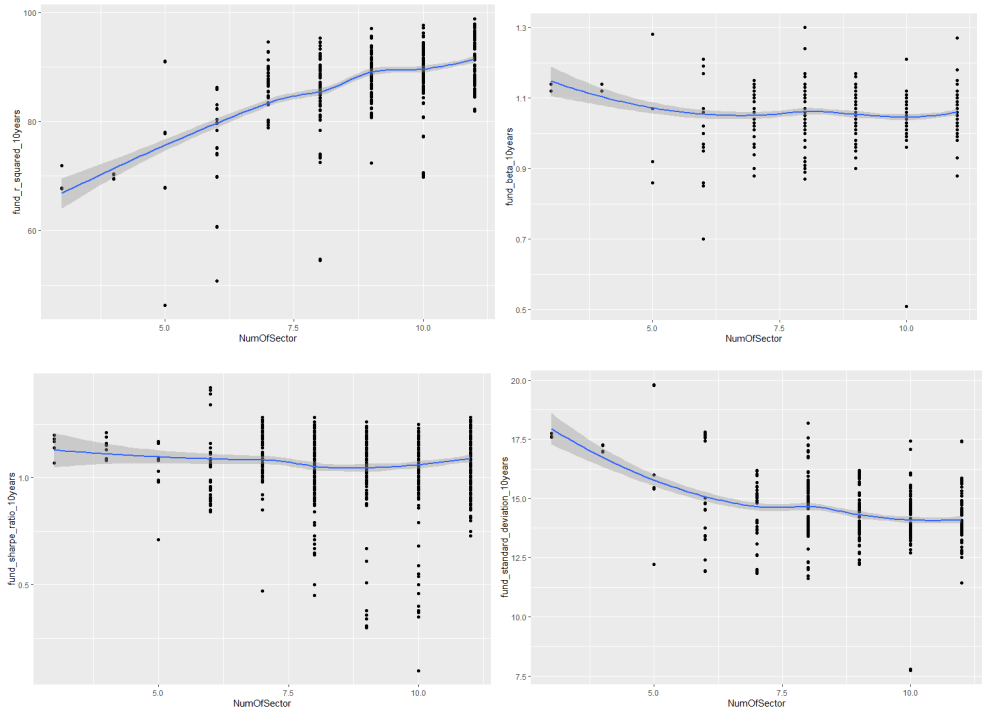
(Figure 3: Rate of Returns (10 years) compared to number of sector with dominant industry)

Number of sectors invested in against 10 year fund r-squared, beta, standard deviation and sharpe ratio

(Starting from top left (Fig 4), in clockwise direction)

Adjusted R-Square increases as the number of sectors increases.

Beta, sharpe ratio and standard deviation decrease and stabilize as the number of sectors increases. This shows it would be better to diversify as the risk factor will go down and stabilize and it would be ideal for managers to give more consistent results.



(Figure 4: Number of sectors compared to R-Square, Beta, Sharpe Ratio, Standard Deviation)

Conclusion:

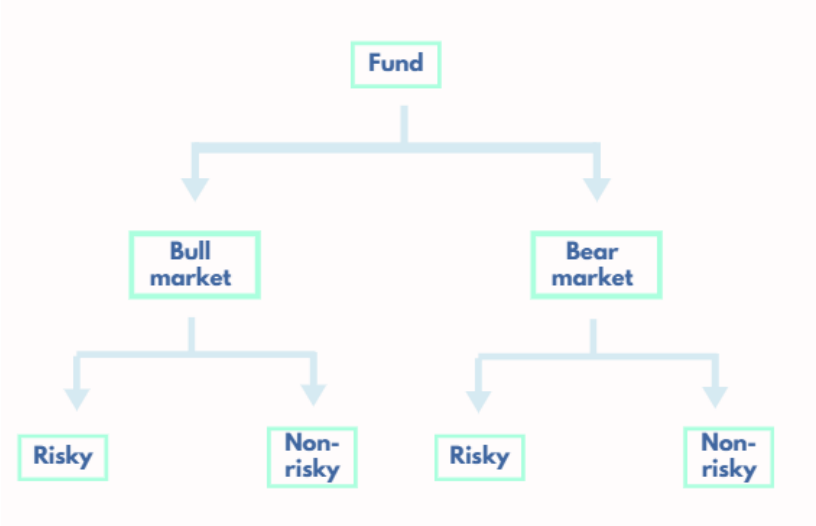
While our initial data exploration shows that higher diversification leads to lower returns, further exploration explains that the results are caused by suboptimal diversification decisions by asset managers. Our findings also concluded that diversification helps in risk mitigation and better performance in the long term. Hence, diversification can be a sustainable strategy that asset managers should be encouraged to adapt and utilise.

4.0 Proposed Solution

4.1 Overall solution:

The report will move on to elaborate on our model building which will eventually form a single model to promote better and more effective diversification methods.

Approach to Solution (Fig 5):



(Figure 5: Overall Framework to Approach)

We will be splitting our data into years that performed well (bull market) and years that did badly (bear market), and then further separate them into risky and non-risky funds.

Due to the nature of our dataset, there will be four different subsets of the dataset that will be studied separately (Good year-risky, Good year-not risky, Bad year-risky, Bad year-not risky). That way, market situation and risk level can be kept constant when running our training models.

1. We first use text mining techniques to do sentiment analysis in order to predict whether the market is going to be bull or bear.
2. Given the different risk levels, Linear Regression models are used to predict the effect of portfolio composition and stock composition on rate of returns of funds.
3. Classification and Regression Tree (CART) model will then be used as a supplement to support and aid asset managers in the usage of the predictive model by linear regression in part (2).

4.1 Text Mining

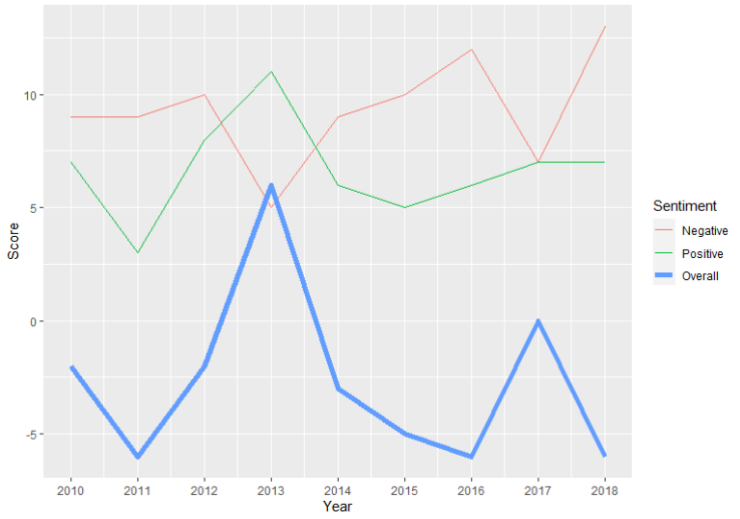
We searched and took the headlines of 15 articles from each year. As the headlines would generally convey the sentiment of the entire article, just having the headlines would be enough to form a trend. After getting the 15 articles for the 8 years, we would read them into our variable and encode them with UTF-8. We would then create a corpus, tokenize the document, removing numbers and punctuation. We would then construct a document feature matrix by using the dfm() function and setting our dictionary using data_dictionary_LSD2015 which is inside the quanteda library before converting the type to dataframe for easier manipulation. We would calculate the adjusted positive, adjusted negative and overall sentiment for each year (Fig 6).

	doc_id	negative	positive	neg_positive	neg_negative	adj.negative	adj.positive	sentiment
1	2010_TextMining.txt	9	7	0	0	9	7	-2
2	2011_TextMining.txt	9	3	0	0	9	3	-6
3	2012_TextMining.txt	10	8	0	0	10	8	-2
4	2013_TextMining.txt	6	10	0	1	5	11	6
5	2014_TextMining.txt	9	6	0	0	9	6	-3
6	2015_TextMining.txt	10	5	0	0	10	5	-5
7	2016_TextMining.txt	12	6	0	0	12	6	-6
8	2017_TextMining.txt	7	7	0	0	7	7	0
9	2018_TextMining.txt	13	7	0	0	13	7	-6

(Figure 6: Sentiment from each each year)

To visualize the sentiment trend, we would plot the adjusted negative, adjusted positive and sentiment on a graph. As seen from the graph, 2010, 2012, 2013 and 2017 have an overall positive or close to neutral score, while 2011, 2015, 2016 and 2018 have a very strong overall negative sentiment (Fig 7).

A research found that bad news dominates the headlines, hence it is possible to have negative news even if the year is a good year. Therefore a low negative sentiment value could still mean a good year. (Stafford, 2014)



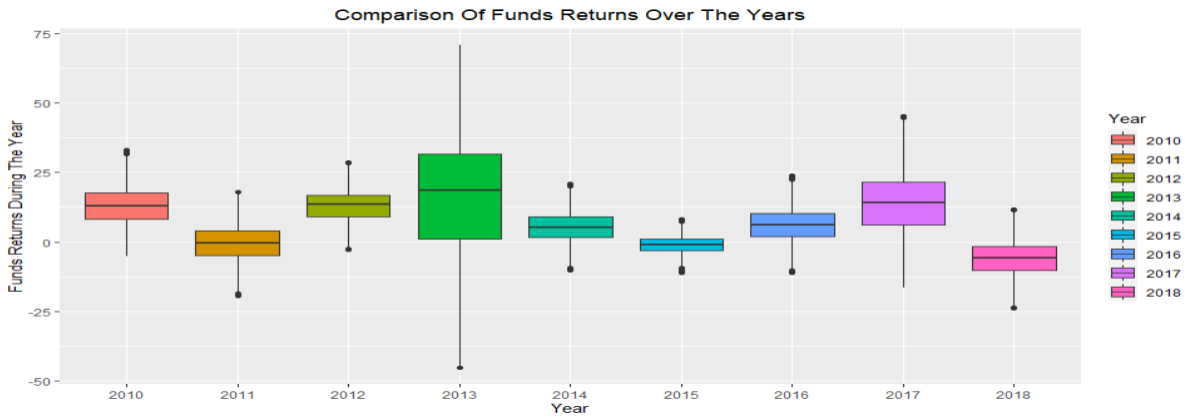
(Figure 7: Sentiment Trend of Positive, Negative and Overall)

We would check if the information from text mining corresponds to the rate of returns. We find the median returns for each of the years by recreating the dataset 9 times for each of the years, removing outliers to remove any extreme returns that may affect the performance for that year. An example is shown from 2018 (Fig 8).

```
delete2018 <- mutualfund[,c("fund_return_2018")]
mutualfund2018 <- mutualfund[complete.cases(delete2018),]
range2018a<- quantile(mutualfund2018$fund_return_2018,0.25)-1.5*IQR(mutualfund2018$fund_return_2018)
range2018b<- quantile(mutualfund2018$fund_return_2018,0.75)+1.5*IQR(mutualfund2018$fund_return_2018)
mutualfund2018 <- mutualfund2018[fund_return_2018 > range2018a & fund_return_2018 < range2018b]
```

(Figure 8: Code to remove NA value and outliers for 2018 fund returns)

We used boxplot to display the returns of 22,000 funds for each year (Fig 9). The bottom 5 medians will be treated as the years where it is the bear market,, while the top 4 medians will be treated as the bull market.



(Figure 9: Boxplot of Fund Returns over the years)

Based on the medians in the graph, we can say that the 2011, 2014, 2015, 2016, 2018 are considered the bear market and 2010, 2012, 2013 and 2017 are considered the bull market.

This is consistent with our results from text mining as the years with an overall strong negative sentiment performed worse than the years with an overall neutral or positive sentiment.

4.2 Linear Regression

Graphical analysis

Based on data exploration, we found that there is a linear relationship between the percentage of sector investment against the average rate of return of a fund.

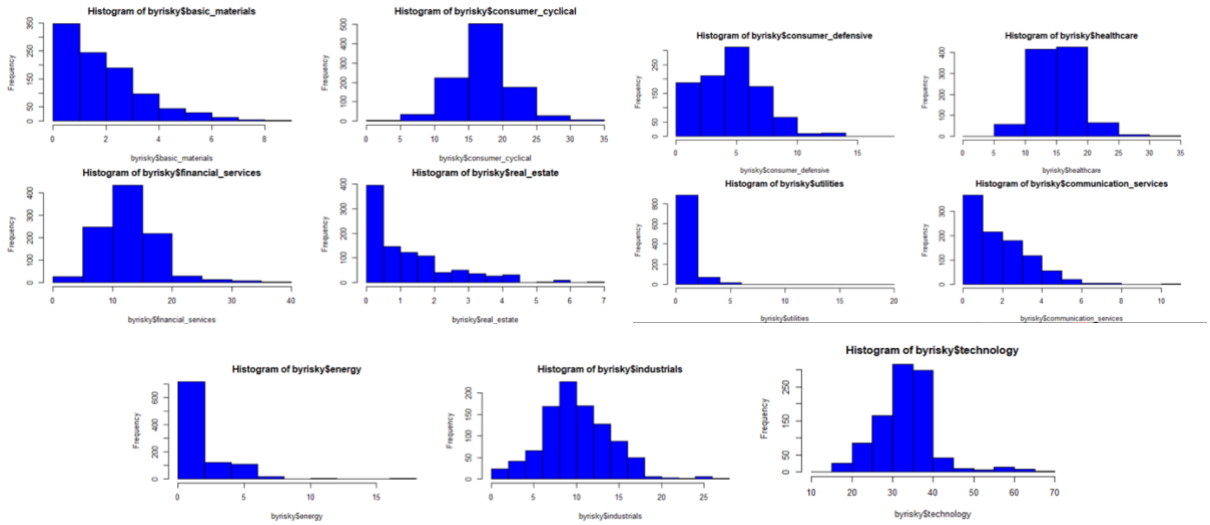
Preparing data for effective model training

Variables from the cleaned dataset comprises of:

- Fund return for individual year
- Number of sectors that a fund invest in
- Sector composition (Industry)
- Portfolio composition (Asset classes)
- Fund alpha and beta
- Derived average rate of return, alpha and beta
- Factor variables of sector composition and portfolio composition (explained below)

A histogram of each sector's percentage shows that there is a specific range for each sector (Fig 10). Thus, using absolute values will contribute to a certain extent of

inaccuracy since the range of values for each sector are inconsistent and that there are some funds with 0% in some of the sectors. We do not want to base our linear regression results on the values which are 0 because a lot of funds invest in 0% of some sectors.



(Figure 10: Histograms of each sector's percentage)

For each individual sector, a new column is created that will reflect the percentage of that sector investment in the fund according to their quantiles with intervals between every 0.1 quantile. Hence there will be 10 levels that reflect the relative percentage. (Appendix C).

By categorising them into factor variables, we found an improvement in the correlation of the individual sector percentage against the average return of the funds (Appendix D). This is further supported by Appendix E where we studied the variance inflation factor (vif) when we run the regression model.

However, categorisation of portfolio composition led to a perfect multicollinearity for linear regression later (Model b) on as the range of each variable is too small (Appendix F).

Modelling process:

For all models, the y variable used will be “avg_return” (average return), a continuous variable. Each subset will undergo stratified train test split through the R package “splitstackshape”, with a split ratio of 70/30.

Linear regression modelling is run on three different combinations of variables to find out which model gives the best accuracy and logical results.

Model a	Y = avg_return X1 = diversification_count X2 = basic_materials X3 = consumer_cyclical X4 financial_services X5 = real_estate X6 = consumer_defensive X7 = healthcare X8 = utilities X9 = communication_services	X10 = energy X11 = industrials X12 = technology X13 = portfolio_cash X14 = portfolio_stocks X15 = portfolio_bonds X16= portfolio_others X17 = portfolio_preferred X18 = portfolio_convertable
Model b	Y = avg_return X1 = diversification_count X2 = basic_materials_level X3 = consumer_cyclical_level X4 financial_services_level X5 = real_estate_level X6 = consumer_defensive_level X7 = healthcare_level X8 = utilities_level X9 = communication_services_level	X10 = energy_level X11 = industrials_level X12 = technology_level X13 = portfolio_cash_level X14 = portfolio_stocks_level X15 = portfolio_bonds_level X16= portfolio_others_level X17 = portfolio_preferred_level X18 = portfolio_convertable_level
Model c	Y = avg_return X1 = diversification_count X2 = basic_materials_level X3 = consumer_cyclical_level X4 financial_services_level	X10 = energy_level X11 = industrials_level X12 = technology_level X13 = portfolio_cash X14 = portfolio_stocks

	X5 = real_estate_level X6 = consumer_defensive_level X7 = healthcare_level X8 = utilities_level X9 = communication_services_level	X15 = portfolio_bonds X16= portfolio_others X17 = portfolio_preferred X18 = portfolio_convertable

Choosing of best model:

Model b is eliminated as the issue of perfect multicollinearity of portfolio composition surfaced. Model c (Fig 11) is picked as the more appropriate model over model a (Appendix G) because the coefficients produced by model c is more logical, where we can deduce the magnitude and direction of the impact for each individual sector.

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.035e+03  9.379e+03   0.110  0.91218
diversification_count       -5.483e-03  1.352e-01  -0.041  0.96768
basic_materials_level       -1.976e-01  7.109e-02  -2.780  0.00560 **
consumer_cyclical_level     -1.964e-01  8.141e-02  -2.412  0.01614 *
financial_services_level    -1.956e-02  8.222e-02  -0.238  0.81199
real_estate_level           -1.101e-01  7.999e-02  -1.376  0.16919
consumer_defensive_level    -1.970e-01  6.927e-02  -2.844  0.00459 **
healthcare_level             8.923e-02  7.353e-02   1.213  0.22542
utilities_level             -1.335e-01  1.821e-01  -0.733  0.46397
communication_services_level -1.756e-01  7.749e-02  -2.266  0.02380 *
energy_level                -1.619e-01  1.037e-01  -1.562  0.11886
industrials_level           -3.229e-02  7.747e-02  -0.417  0.67696
technology_level             1.411e-01  1.020e-01   1.384  0.16678
portfolio_cash              -1.002e+01  9.378e+01  -0.107  0.91495
portfolio_stocks            -1.004e+01  9.378e+01  -0.107  0.91478
portfolio_bonds             -9.420e+00  9.378e+01  -0.100  0.92002
portfolio_others            -9.351e+00  9.376e+01  -0.100  0.92059
portfolio_preferred         -8.929e+00  9.380e+01  -0.095  0.92419
portfolio_convertable       -1.232e+01  9.480e+01  -0.130  0.89660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.69 on 635 degrees of freedom
Multiple R-squared:  0.3842,    Adjusted R-squared:  0.3668
F-statistic: 22.01 on 18 and 635 DF,  p-value: < 2.2e-16
```

(Figure 11: Model c linear regression summary)

The linear regression model c has an adjusted r-square of 0.3668 and coefficients that can be easily interpreted by asset managers. Root mean squared error is at 4.49% (Fig 12) with the mean “avg_return” of the dataset being +24%.

```
> rmse2c_test
[1] 4.490916
```

(Figure 12: Root mean squared error of model c)

Insights and analysis:

The adjusted r-square and root mean squared error suggests that the regression model may not be the most accurate, however this is just step one of the optimisation, and some limitations are discussed in section 5.

With the coefficients from the regression model, asset managers are able to predict the effects of different sector stocks on their funds, allowing them to make necessary adjustments to optimise the rate of return of their funds.

However, to get the optimal range of each sector composition, CART analysis will come into play, which will be discussed in the section below.

4.3 CART Modelling

Purpose:

The ***rpart*** package was used to develop the CART model. CART, a type of decision tree, was selected as it has high explainability and ensures that the model is easy to understand and fully transparent. Managers view this model can better appreciate the decision-making process which can act as a guideline for their portfolio construction (explained below) and is more effective in convincing them to make changes to their portfolio to overcome their biases.

Dataset preparation:

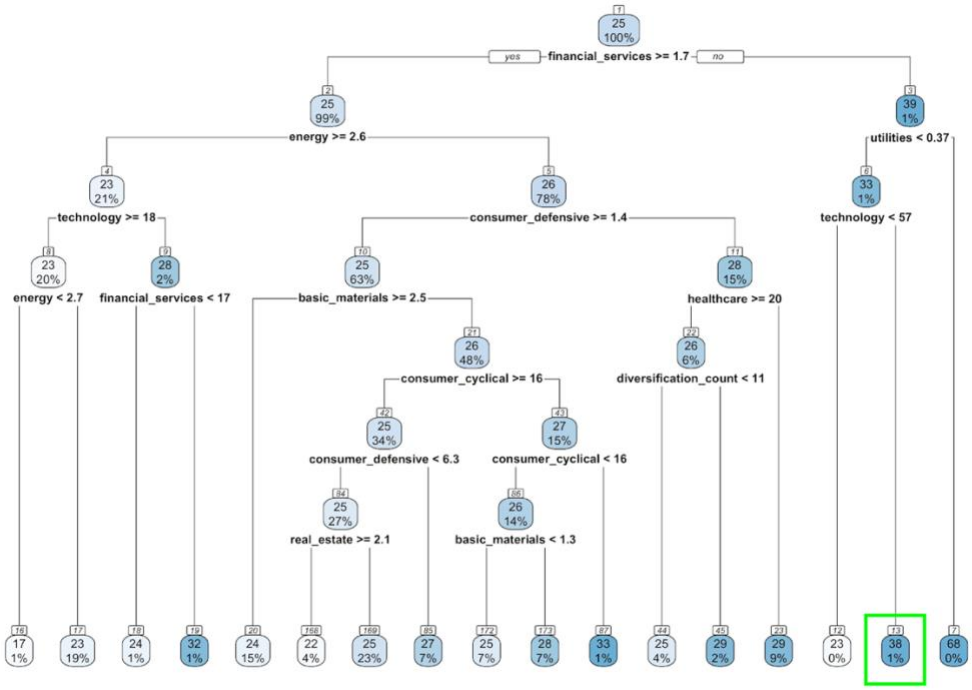
Variables from the cleaned dataset comprises of:

- Derived average rate of return (of Good and Bad years)
- Number of sectors that a fund invest in (for Industry sector analysis)

- Industry sectors composition (for Industry sector analysis)
- Portfolio (Asset classes) composition (for Asset classes analysis)

Modelling process:

To prevent the issue of overfitting where testset error increases alongside model complexity and the issue of the tree being unstable and too sensitive to small changes in the data, the tree will be pruned to obtain the model where the error is within one standard error of the best model such that the trees are statistically equivalent in terms of error.



(Figure 13 : CART model (Good Year Risky) for Industry Sectors)

The above model (Fig 13) is developed for the analysis of the industry sectors composition on fund returns on a good performing year . Models for the 7 remaining scenarios will be included in Appendix H along with its relevant explanations. From Figure 13 , it can be observed that Terminal Node 7 yields an average return of 68%. However, this number is inflated with only 2 data within and is likely to be an outlier. Hence, we decided to use the next optimal tree path, Terminal Node 13, which yields an average return of 38% which is within the reasonable limits. The optimal tree path is as follows: *financial services<1.7, utilities<0.37, technology≥57*. This results matches our findings from data exploration where returns increase with greater

weightage in the technology sector, and returns are higher when there is a lower proportion of the funds invested in financial services and utilities. The above model has a cross validation mean squared error of 18. An error of around 4.2 is within the acceptable limits given that the median average return was 24.5.

The figure (Fig 14) shows the importance of each variable when developing the decision tree. Financial services, consumer defensive and technology are some of the more important variables that require more attention from asset managers.

```
> gyricart2$variable.importance #Decreasing order of importance: financial_services, consumer_defensive, technology
    financial_services     consumer_defensive             technology             healthcare              utilities  diversification_count
             5182.9573              4056.8552              3606.4209              2881.6941              2741.0877              2306.8015
                energy            industrials      consumer_cyclical       basic_materials          real_estate communication_services
             1559.0588              1174.0825               850.1491               828.8906               770.3363               254.8890
```

(Figure 14: Variable importance for Industry Sectors (Good Year Risky))

Link to linear regression model:

The optimal tree path obtained from this CART model is as follows: *financial services<1.7, utilities<0.37, technology≥57*.

This range serves as a guideline for asset managers to use with linear regression coefficients. (Explained below in “Conclusion - Interface”)

5.0 Limitations and Model Improvements

5.1 Limitations

There are some limitations faced when using these models. The rapidly changing market suggests that historical data may not be able to predict returns in future years as accurately. For instance, the industrial sector used to be profitable 3 years ago but led to losses instead in recent years (Fidelity, 2020). Therefore, up-to-date data must be used in developing our models to remain relevant. Since White Rock mainly engages in active management of funds, a possible method would be to use quarterly or monthly data to train the model so that the results are reflective of the current market.

The US mutual funds were chosen because they have the largest value of assets in mutual funds (Statista, 2020). As the dataset only comes from the US itself, it does not include any overseas funds to provide a global sense of the market. Each country will have different types of regulations to adhere by and this will influence the composition of the funds. Moreover, this dataset can only be applied to US citizens since only they are eligible to invest in US registered funds (Brown, 2019). Therefore, new datasets need to be obtained and trained for different countries to give a different composition for that country's citizens.

Clients' risk appetite are only categorised into 2 categories: aggressive investors (beta >1) and risk averse investors (0 <beta ≤1). This may be an oversimplification of risk tolerance level as there is a spectrum of risk level different clients would be willing to take. Hence, a solution is to split beta into different bins to consider the different risk levels.

5.2 Project Extension

Beside manually going and picking the top 15 articles, we can get a bot to web scrape each year and give us the market sentiment for each year. This would be more conclusive as there would be more results to gauge the sentiment of the market. This enables the team to predict with greater certainty if the market would be good or bad. To be more catered to White Rock, a daily or monthly web scraping can be used to judge the market sentiment. This would enable managers to predict with greater certainty how the market would react for a shorter period of time.

Linear regression may not be most ideal in predicting the fund returns as they do not follow a linear relationship which generates a low R-squared as a result. Other models including Random Forest Regression can help us in getting a better model as it would be able to produce a classifier that is more accurate (Bakshi, 2020). This is due to random forest capability in running numerous decision trees and combining them to an average one.

Moving Average Convergence Divergence (MACD) can also be used to complement our analysis and provide new insights. MACD is useful in providing an early indication of a trend reversal (Investopedia, 2020). This would enable White Rock to

identify the momentum of the market, if it is good or bad, to allow them to have a better idea of when to buy and sell stocks (Hayes, 2020)

6.0 Conclusion

Interface: “Bye Bye-as (Bi-as)’

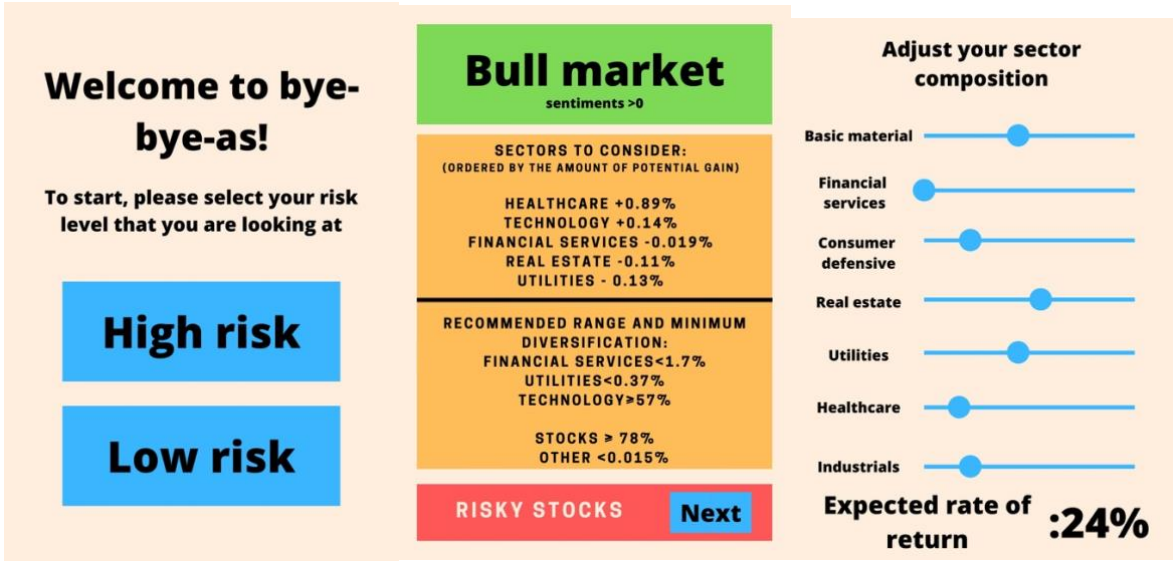
An interactive mobile application interface will be developed using results from the linear regression and CART models.

When the manager accesses this application, the application will evaluate the current year market sentiment and would prompt the manager to select the risk level of their clients (Risky/ Not Risky).

They will be shown the recommended composition computed by linear regression coefficients and CART model optimal tree path. Coefficients from linear regression model gives an idea of how will a particular sector affect the rate of return while the combination output from CART gives a guide to the range of percentages that the asset managers should use for the recommended sectors.

The output range from CART results are very useful because it will ensure that a suitable amount of diversification. For example, if the asset manager only follows the linear regression result, due to the lack of information, he will just pick the top performing sector and invest 100% in it. Hence CART results act as a supplement for easier interpretation of the linear regression model.

From there, an additional feature would allow the manager to switch between the generated optimal path and a console for his own input. This console will generate the approximate returns when the manager adjusts the industry and asset class weightage.


Welcome to bye-bye-as!
To start, please select your risk level that you are looking at
High risk
Low risk
Bull market
sentiments >0
SECTORS TO CONSIDER:
(ORDERED BY THE AMOUNT OF POTENTIAL GAIN)
HEALTHCARE +0.89%
TECHNOLOGY +0.14%
FINANCIAL SERVICES -0.019%
REAL ESTATE -0.11%
UTILITIES - 0.13%
RECOMMENDED RANGE AND MINIMUM DIVERSIFICATION:
FINANCIAL SERVICES<1.7%
UTILITIES<0.37%
TECHNOLOGY>57%
STOCKS > 78%
OTHER <0.015%
RISKY STOCKS
Next
Adjust your sector composition
Basic material
Financial services
Consumer defensive
Real estate
Utilities
Healthcare
Industrials
Expected rate of return
:24%

7.0 Appendix

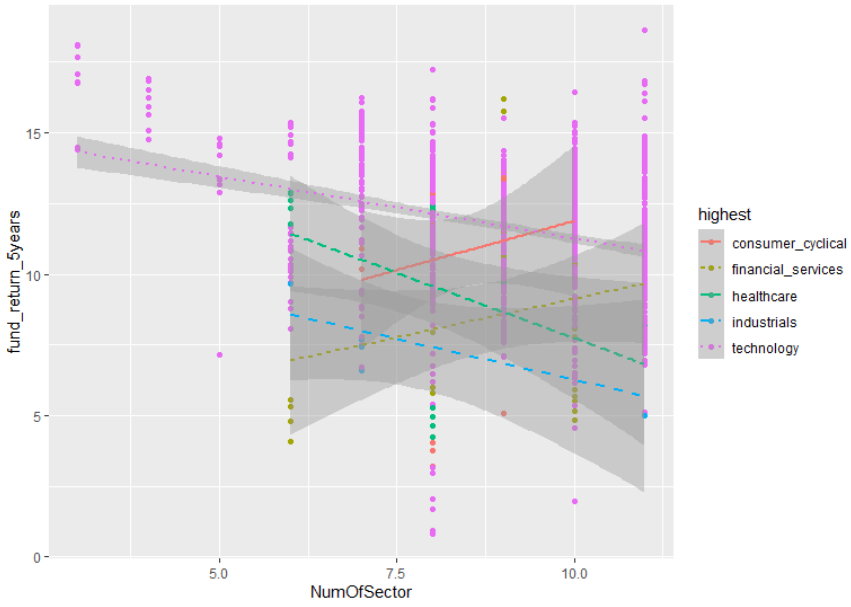
Appendix A

Type of Column	*Columns Removed*
Bond Related Columns	*bond_maturity, bond_duration, rating_us_government, rating_aaa, rating_aa, rating_a, rating_bbb, rating_bb, rating_b, rating_below_b, rating_other*
Category Related Columns	*net_annual_expense_ratio_category, category_return_ytd, category_return_1month, category_return_3months, category_return_1year, category_return_3years, category_return_5years, category_return_10years, category_return_2018, category_return_2017, category_return_2016, category_return_2015, category_return_2014, category_return_2013, category_return_2012, category_return_2011, category_return_2010, category_alpha_3years, category_alpha_5years, category_alpha_10years, category_beta_3years, category_beta_5years, category_beta_10years, category_mean_annual_return_3years, category_mean_annual_return_5years, category_mean_annual_return_10years, category_r_squared_3years, category_r_squared_5years, category_r_squared_10years, category_standard_deviation_3years, category_standard_deviation_5years, category_standard_deviation_10years, category_sharpe_ratio_3years, category_sharpe_ratio_5years, category_sharpe_ratio_10years*

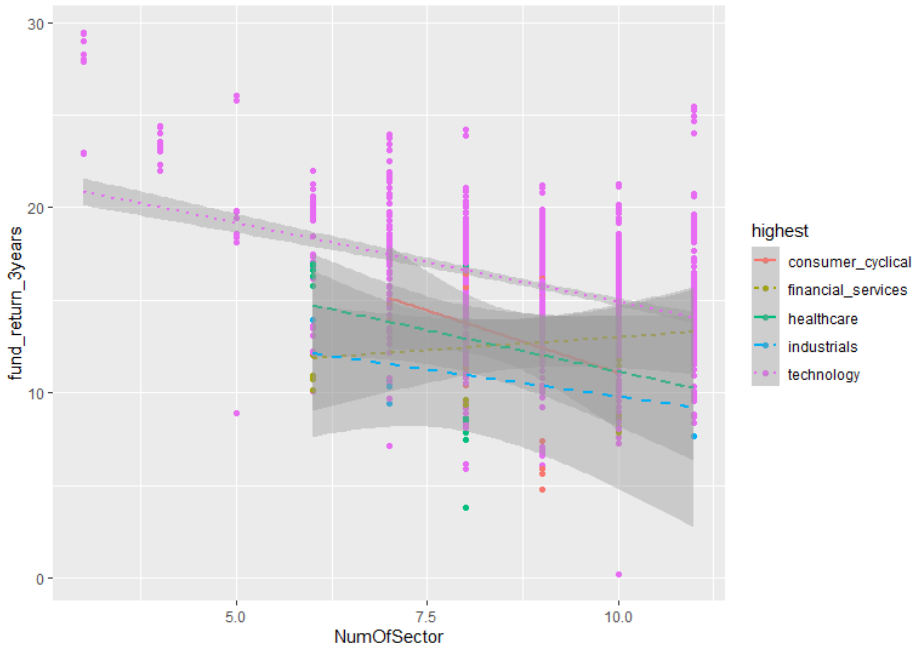
MorningStar Rating Columns	*morningstar_rating, morningstar_return_rating, morningstar_risk_rating*
Price Ratio Columns	*price_earnings, price_book, price_sales, price_cashflow*

Appendix B

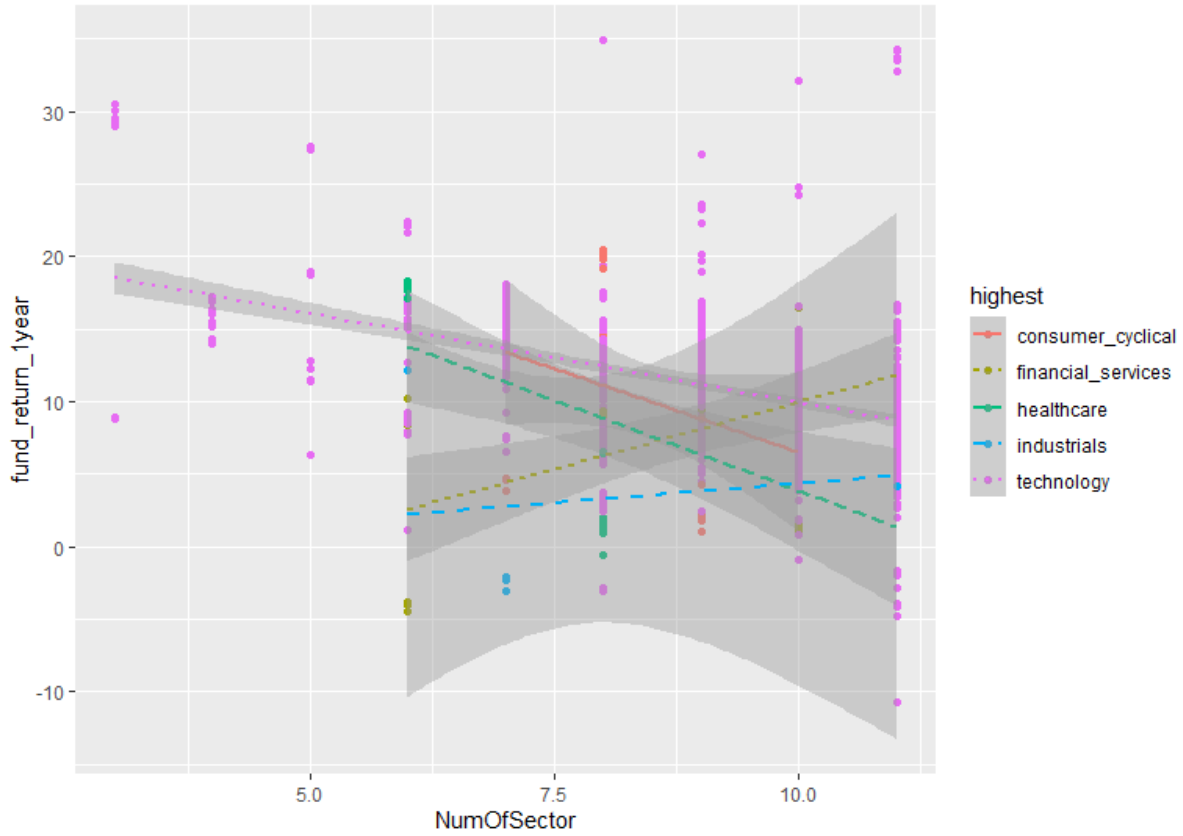
5 years



3 years



1 year



Appendix C

Factoring of data (LR)

0 to 0.1 Quantile: Level 1	0 to 0.1 Quantile: Level 1
0.1 to 0.2 Quantile: Level 2	0.1 to 0.2 Quantile: Level 2
0.2 to 0.3 Quantile: Level 3	0.2 to 0.3 Quantile: Level 3
0.3 to 0.4 Quantile: Level 4	0.3 to 0.4 Quantile: Level 4
0.4 to 0.5 Quantile: Level 5	0.4 to 0.5 Quantile: Level 5
0.5 to 0.6 Quantile: Level 6	0.5 to 0.6 Quantile: Level 6
0.6 to 0.7 Quantile: Level 7	0.6 to 0.7 Quantile: Level 7
0.7 to 0.8 Quantile: Level 8	0.7 to 0.8 Quantile: Level 8
0.8 to 0.9 Quantile: Level 9	0.8 to 0.9 Quantile: Level 9
0.9 to 1 Quantile: Level 10	0.9 to 1 Quantile: Level 10

Appendix D

Improvement in correlation for sector composition

basic_materials	0.031693687
consumer_cyclical	-0.004168620
financial_services	-0.024778828
real_estate	0.214860045
consumer_defensive	-0.190009570
healthcare	-0.094102491
utilities	-0.119619062
communication_services	-0.100810572
energy	0.008867622
industrials	-0.197185224
technology	0.250959414
basic_materials_level	0.112476820
consumer_cyclical_level	-0.030872202
financial_services_level	-0.097489880
real_estate_level	-0.038011377
consumer_defensive_level	-0.225460356
healthcare_level	0.027416934
utilities_level	-0.084085605
communication_services_level	-0.088890365
energy_level	-0.048480084
industrials_level	-0.161249751
technology_level	0.240455862

Appendix E

Further support for correlation from vif

VIF for Model 2a

```
> vif(model2a)
 diversification_count        basic_materials      consumer_cyclical      financial_services           real_estate      consumer_defensive
          1.335163e+00           3.591382e+04           2.231443e+05           3.319244e+05           2.671826e+04           1.013176e+05
           healthcare              utilities communication_services                 energy            industrials             technology
          1.877531e+05           4.054549e+04           4.309529e+04           6.406894e+04           2.397246e+05           8.015075e+05
       portfolio_cash       portfolio_stocks        portfolio_bonds       portfolio_others    portfolio_preferred portfolio_convertable
          3.413441e+06           1.542388e+07           7.029594e+06           2.573496e+05           1.780072e+05           1.910141e+03
```

VIF for Model 2b

```
> vif(model2b)
Error in vif.default(model2b) :
  there are aliased coefficients in the model
```

VIF for Model 2c

```
> vif(model2c)
       diversification_count        basic_materials_level      consumer_cyclical_level     financial_services_level
                1.928191e+00                 1.258723e+00                 2.269615e+00                 2.304425e+00
          real_estate_level     consumer_defensive_level           healthcare_level              utilities_level
                1.407217e+00                 1.462496e+00                 1.875821e+00                 1.725273e+00
communication_services_level                energy_level           industrials_level            technology_level
                1.632535e+00                 2.021245e+00                 2.266143e+00                 3.524080e+00
               portfolio_cash             portfolio_stocks              portfolio_bonds             portfolio_others
                3.417110e+06                 1.544066e+07                 7.037089e+06                 2.576149e+05
         portfolio_preferred       portfolio_convertable
                1.781946e+05                 1.915254e+03
```

Appendix F

Portfolio composition factoring

```
portfolio_cash                0.244095736
portfolio_stocks             -0.301633031
portfolio_bonds               0.143203761
portfolio_others              0.255006336
portfolio_preferred           0.186610018
portfolio_convertable         0.152587544
portfolio_cash_level          0.261909322
portfolio_stocks_level       -0.254266220
portfolio_bonds_level                  NA
portfolio_others_level        0.218772463
portfolio_preferred_level     0.028718843
portfolio_convertable_level            NA
```

Appendix G

Summary for Model 2a

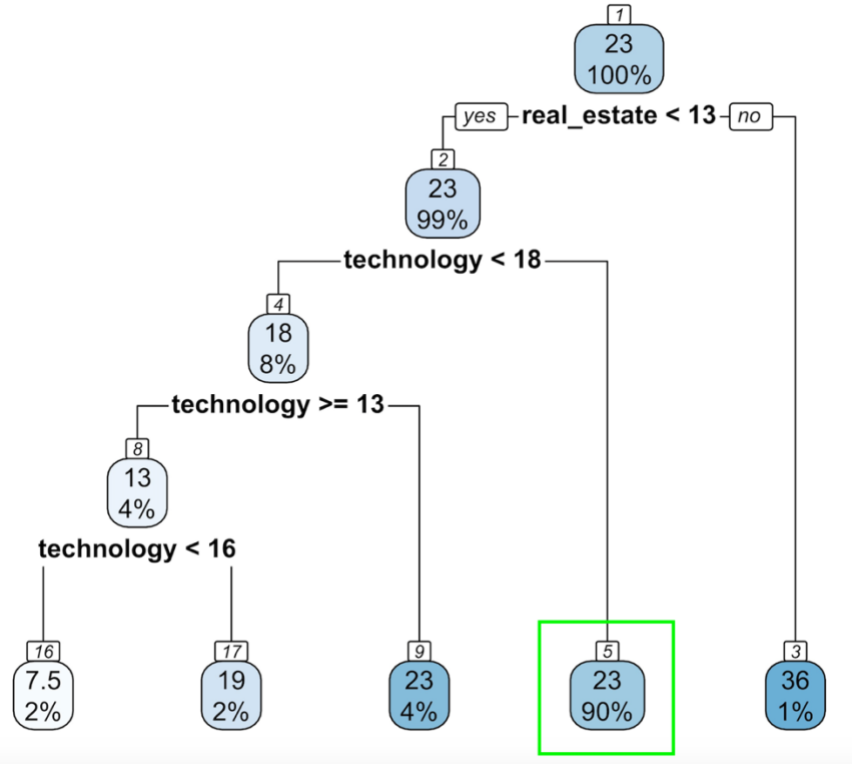
```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          9262.4945  9394.1236   0.986    0.325
diversification_count   -0.0974     0.1112  -0.876    0.381
basic_materials       -23.9072    17.1445  -1.394    0.164
consumer_cyclical     -23.8265    17.1548  -1.389    0.165
financial_services    -23.6856    17.1545  -1.381    0.168
real_estate           -23.7317    17.1594  -1.383    0.167
consumer_defensive    -23.8912    17.1517  -1.393    0.164
healthcare            -23.6495    17.1509  -1.379    0.168
utilities             -23.8354    17.1664  -1.388    0.165
communication_services -23.7867   17.1612  -1.386    0.166
energy                -23.8726    17.1488  -1.392    0.164
industrials           -23.7035    17.1515  -1.382    0.167
technology            -23.6087    17.1534  -1.376    0.169
portfolio_cash        -68.6628    92.5986  -0.742    0.459
portfolio_stocks      -68.6677    92.5985  -0.742    0.459
portfolio_bonds       -68.0910    92.5994  -0.735    0.462
portfolio_others      -68.0120    92.5802  -0.735    0.463
portfolio_preferred   -67.4808    92.6155  -0.729    0.467
portfolio_convertable -69.1068    93.5301  -0.739    0.460

Residual standard error: 3.646 on 635 degrees of freedom
Multiple R-squared:  0.399,     Adjusted R-squared:  0.382
F-statistic: 23.42 on 18 and 635 DF,  p-value: < 2.2e-16
```

Appendix H

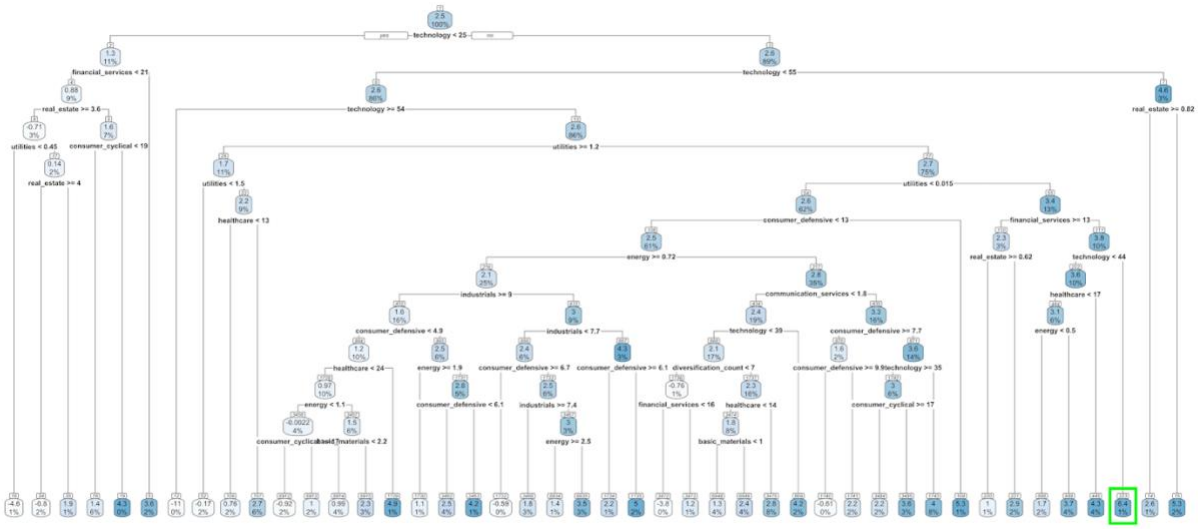
CART model (Good Year Not Risky) for Industry Sectors

Optimal Tree Path: Real estate <13, Technology >= 18



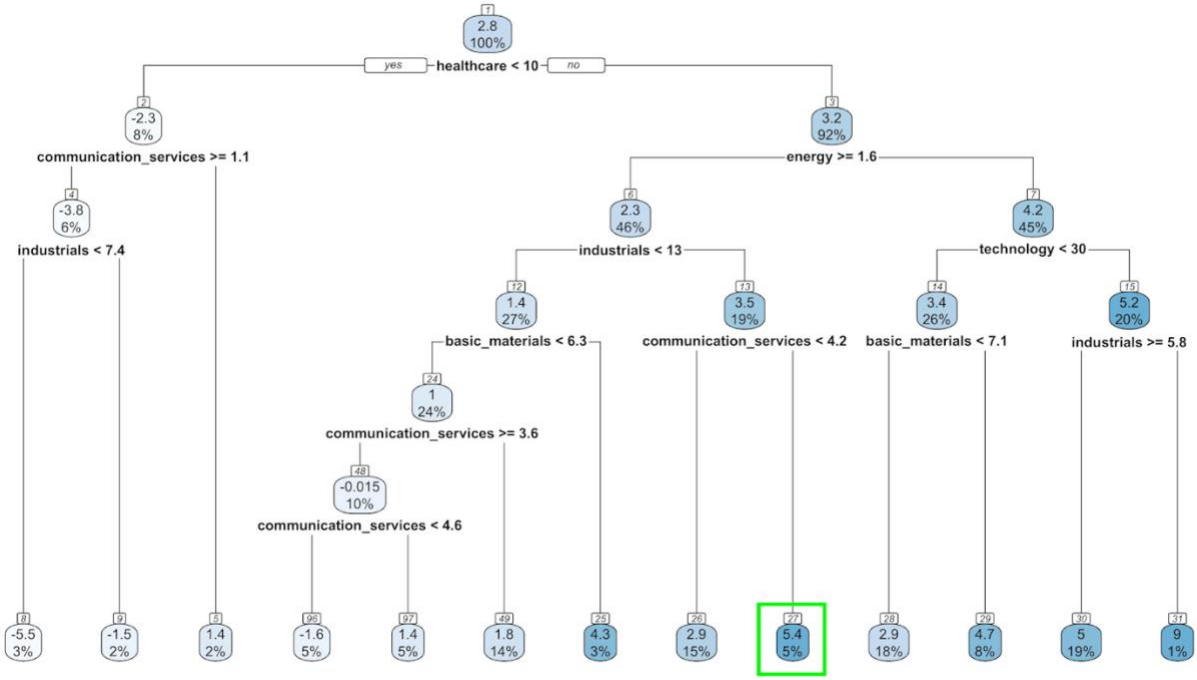
CART model (Bad Year Risky) for Industry Sectors

Optimal Tree Path: 44<=Technology<54, 0.015<=Utilities<1.2, Financial_services<13



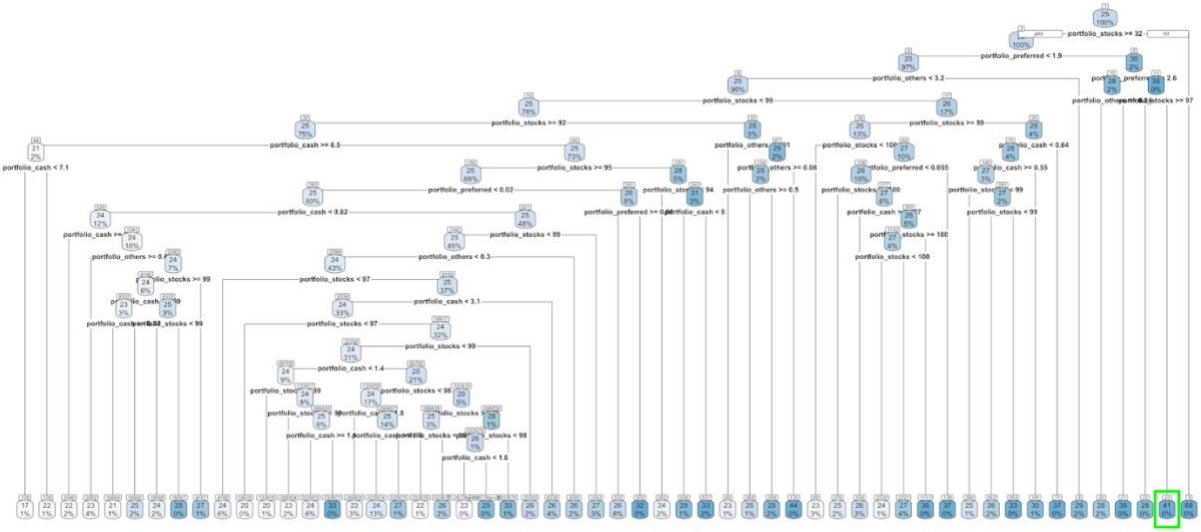
CART model (Bad Year Not Risky) for Industry Sectors

Optimal Tree Path: Healthcare>=10, energy>=1.6, industrials>=13, Communication_services>=4.2



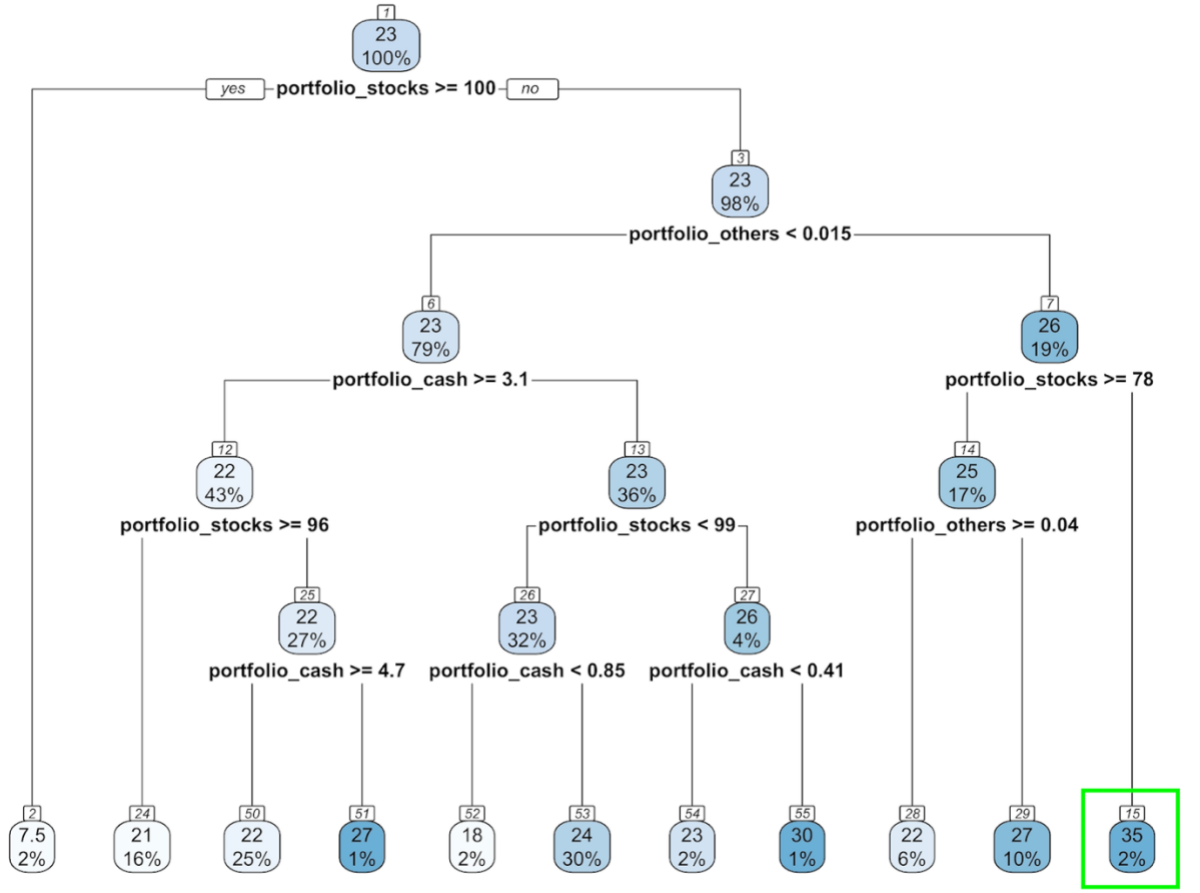
CART model (Good Year Risky) for Asset Classes

Optimal Tree Path: 32<=Portfolio_stocks<97, 1.9<=Portfolio_preferred<2.6



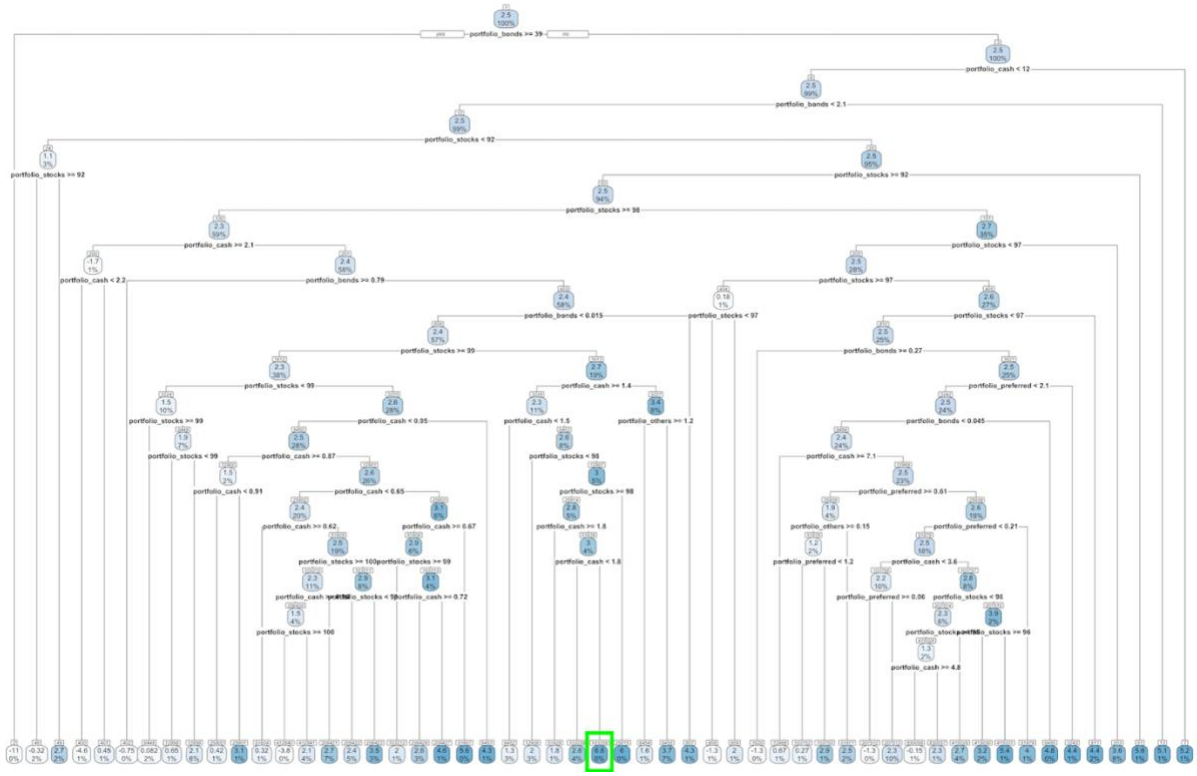
CART model (Good Year Not Risky) for Asset Classes

Optimal Tree Path: Portfolio_others>=0.015, Portfolio_stocks<78



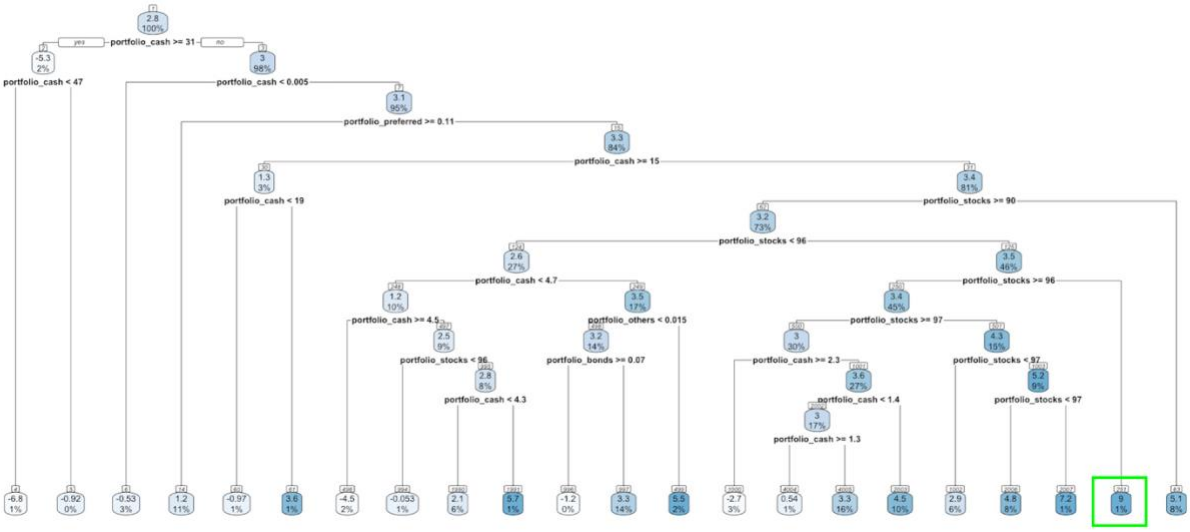
CART model (Bad Year Risky) for Asset Classes

Optimal Tree Path: Portfolio_bonds <0.015, 98<=Portfolio_stocks<99, 1.5<=Portfolio_cash<=1.8 (Portfolio_cash~1.8)



CART model (Bad Year Not Risky) for Asset Classes

Optimal Tree Path: 0.005<=Portfolio_cash<15, Portfolio_preferred<0.11, Portfolio_stocks~96



8.0 References

Bakshi, C. (2020, June 09). Random Forest Regression. Retrieved November 01, 2020, from https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Brown, M. (2019, June 25). How Mutual Funds Differ Around the World. Retrieved 1 November 2020, from https://www.investopedia.com/articles/mutualfund/08/foreign-mutual-funds.asp

Hayes, A. (2020, October 07). Moving Average Convergence Divergence (MACD) Definition. Retrieved November 01, 2020, from https://www.investopedia.com/terms/m/macd.asp

Hoffman, N., Huber, M., & Smith, M. (2017, December 19). An analytics approach to debiasing asset-management decisions. Retrieved October 31, 2020, from https://www.mckinsey.com/industries/financial-services/our-insights/an-analytics-approach-to-debiasing-asset-management-decisions

Investopedia. (2020, April 3). What are the best technical indicators to complement the moving average convergence divergence (MACD)? Retrieved November 01, 2020, fromhttps://www.investopedia.com/ask/answers/122314/what-are-best-technical-indicators-complement-moving-average-convergence-divergence-macd.asp

Kase Fund Manager – Fixed Income and Multi-Asset See all articles, M. (2018, November 24). The risk of familiarity bias in asset allocation. Retrieved October 31, 2020, from https://www.schroders.com/en/au/advisers/insights/the-fix/the-risk-of-familiarity-bias-in-asset-allocation/

Kenton, W. (2020, July 28). Behavioral Finance Definition. Retrieved October 31, 2020, from https://www.investopedia.com/terms/b/behavioralfinance.asp

Kerrigan, M., Williams, D., Smith, K., Petitto, J., & Nolting, D. (2020, July 13). Power of Data-driven Asset Management. Retrieved November 01, 2020, from

https://www.accenture.com/us-en/insights/capital-markets/power-data-driven-asset-management

Mutual fund assets by country 2019 | Statista. (2020, February). Retrieved 1 November 2020, from https://www.statista.com/statistics/270289/amount-of-fund-assets-in-selected-countries-of-the-world/

Mutual fund outlook: The time to act is now. (2019, July). Retrieved 1 November 2020, from https://www.pwc.com/us/en/industries/financial-services/library/mutual-fund-outlook.html

Nath, T. (2019, June 25). How Big Data Has Changed Finance. Retrieved October 31, 2020, from https://www.investopedia.com/articles/active-trading/040915/how-big-data-has-changed-finance.asp

Sectors & Industries - Performance - Fidelity. (2020, October 30). Retrieved 1 November 2020, from https://eresearch.fidelity.com/eresearch/markets_sectors/sectors/si_performance.jhtml?tab=siperformance

Stafford, T. (2014). Psychology: Why bad news dominates the headlines. Retrieved November 01, 2020, from https://www.bbc.com/future/article/20140728-why-is-all-the-news-bad

Waggoner, J. (2019, October 21). The 25 Best Mutual Funds of All Time. Retrieved October 31, 2020, from https://www.kiplinger.com/slideshow/investing/t041-s001-the-25-best-mutual-funds-of-all-time/index.html