Department of Electrical Engineering

FINAL YEAR PROJECT REPORT

BENGEGU4-CDE-2019/20-CWT-01

Project Title:
Portfolio Management with Heuristic Optimization

Student Name: YOGISH Suhas
Student ID: 54482837
Supervisor: Mr TING, Van C W
Assessor: Mr NG, K T

Bachelor of Engineering (Honours) in
Computer and Data Engineering

Student Final Year Project Declaration

I have read the student handbook and I understand the meaning of academic dishonesty, in particular plagiarism and collusion. I declare that the work submitted for the final year project does not involve academic dishonesty. I give permission for my final year project work to be electronically scanned and if found to involve academic dishonesty, I am aware of the consequences as stated in the Student Handbook.

Project Title : Portfolio Management with Heuristic Optimization

Student Name: YOGISH Suhas

Student ID: 54482837

Signature: Suhas Yogish

Date: 20/04/2020

Turnitin Originality Report

Processed on: 20-Apr-2020 18:20 HKT
ID: 1301304548
Word Count: 8189
Submitted: 3

FYP54482837.pdf By Suhas YOGISH

Similarity Index	Similarity by Source
14%	Internet Sources: 6% Publications: 3% Student Papers: 13%

1% match (student papers from 10-Apr-2017)
Submitted to City University of Hong Kong on 2017-04-10

1% match (student papers from 10-Apr-2019)
Submitted to City University of Hong Kong on 2019-04-10

1% match (Internet from 06-Nov-2018)
https://www.quantopian.com/help

< 1% match (student papers from 04-Feb-2019)
Submitted to University of Portsmouth on 2019-02-04

< 1% match (student papers from 11-Dec-2014)
Submitted to University of Arizona on 2014-12-11

< 1% match (student papers from 02-Mar-2020)
Submitted to University of Birmingham on 2020-03-02

< 1% match (student papers from 31-May-2019)
Submitted to Kingston University on 2019-05-31

< 1% match (Internet from 15-Jan-2020)
https://scitepress.org/Papers/2019/83465/pdf/index.html

< 1% match (publications)
Prabas Gayadeeptha, Tharindu Pushpitha Baddewithana, Kalpana Vijintha Pannegama, Chamari Sandarekha Samarakkody et al. "EyeDriver: Intelligent Driver Assistance System", 2019 14th Conference on Industrial and Information Systems (ICIIS), 2019

< 1% match (student papers from 07-Nov-2018)
Submitted to Universidad EAFIT on 2018-11-07

< 1% match (student papers from 09-May-2019)
Submitted to University of Greenwich on 2019-05-09

< 1% match (student papers from 22-Nov-2011)
Submitted to University of St. Gallen on 2011-11-22

< 1% match (Internet from 08-Jun-2019)
https://medium.com/greyatom/boosting-ce84639a805d

< 1% match (student papers from 16-Aug-2013)
Submitted to University of Edinburgh on 2013-08-16

< 1% match (student papers from 11-Mar-2020)
Submitted to Hult International Business School on 2020-03-11

< 1% match (student papers from 20-Mar-2014)
Submitted to Brunel University on 2014-03-20

< 1% match (student papers from 10-Nov-2019)
Submitted to Macquarie University on 2019-11-10

< 1% match (student papers from 21-Aug-2019)
Submitted to University of Portsmouth on 2019-08-21

< 1% match (student papers from 14-Dec-2019)
Submitted to Lynn University on 2019-12-14

< 1% match (student papers from 10-Jan-2013)
Submitted to University of Southampton on 2013-01-10

< 1% match ()
http://researchrepository.murdoch.edu.au/id/eprint/5817/

< 1% match (publications)
Haiying Jia, Roar Adland, Yuchen Wang. "Latin American Oil Export Destination Choice: A Machine Learning Approach", 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2019

< 1% match (Internet from 30-Mar-2010)
http://www.allmericafinancial.com/pdf/suplmnts/funds/AIMpefS1.pdf

< 1% match (student papers from 08-Oct-2017)
Submitted to University of Newcastle on 2017-10-08

< 1% match (Internet from 31-Jan-2020)
https://res.mdpi.com/d_attachment/sustainability/sustainability-12-00318/article_deploy/sustainability-12-00318.pdf

< 1% match (publications)
"Intelligent Computing", Springer Science and Business Media LLC, 2019

< 1% match (student papers from 24-Sep-2019)
Submitted to University of Westminster on 2019-09-24

< 1% match (student papers from 31-Mar-2020)
Submitted to University of Computer Studies on 2020-03-31

< 1% match (student papers from 20-Jan-2013)
Submitted to University of Northumbria at Newcastle on 2013-01-20

< 1% match (Internet from 08-Aug-2019)
https://savart.in/blog/modern-portfolio-theory-a-complete-guide/

< 1% match (student papers from 18-May-2017)
Submitted to London Business School on 2017-05-18

< 1% match (Internet from 12-Dec-2019)
https://www.tradingsig.com/chicago-mercantile-exchange.html

< 1% match (student papers from 09-Oct-2016)
Submitted to Kennesaw State University on 2016-10-09

< 1% match ()
http://theses.gla.ac.uk/2110/

< 1% match (student papers from 06-Nov-2018)
Submitted to Stevens Institute of Technology on 2018-11-06

< 1% match (student papers from 16-Mar-2018)
Submitted to University of Durham on 2018-03-16

< 1% match (student papers from 14-Nov-2019)
Submitted to Syiah Kuala University on 2019-11-14

< 1% match (student papers from 25-Nov-2015)
Submitted to Trinity College Dublin on 2015-11-25

< 1% match (student papers from 17-May-2019)
Submitted to University of Queensland on 2019-05-17

< 1% match (student papers from 21-Aug-2019)
Submitted to University of Portsmouth on 2019-08-21

< 1% match (student papers from 02-May-2016)
Submitted to Georgetown University on 2016-05-02

< 1% match (student papers from 22-Oct-2018)
Submitted to Higher Education Commission Pakistan on 2018-10-22

< 1% match (Internet from 07-Jun-2017)

http://lbms03.cityu.edu.hk/studproj/ee/2007eeykw955.pdf

< 1% match (student papers from 10-Apr-2019)
Submitted to City University of Hong Kong on 2019-04-10

< 1% match (student papers from 29-Apr-2019)
Submitted to University of Reading on 2019-04-29

< 1% match (student papers from 02-May-2016)
Submitted to University of Warwick on 2016-05-02

< 1% match (student papers from 19-Sep-2016)
Submitted to Emirates Aviation College, Aerospace & Academic Studies on 2016-09-19

< 1% match (student papers from 26-Mar-2009)
Submitted to University of Lancaster on 2009-03-26

< 1% match (student papers from 11-Oct-2011)
Submitted to University of Stellenbosch, South Africa on 2011-10-11

< 1% match (student papers from 05-Oct-2016)
Submitted to Imperial College of Science, Technology and Medicine on 2016-10-05

< 1% match (student papers from 13-Sep-2018)
Submitted to University of Warwick on 2018-09-13

< 1% match (Internet from 15-Mar-2020)
https://myassignmenthelp.com/free-samples/risk-and-social-theory-in-environmental-management?access-library-email=

< 1% match (Internet from 16-Dec-2019)
https://lup.lub.lu.se/luur/download?fileOId=4499311&func=downloadFile&recordOId=4499301

< 1% match (publications)
Chi Cheong Allen Ng, Jianfu Shen. "Screen winners from losers using simple fundamental analysis in the Pacific-Basin stock markets", Pacific-Basin Finance Journal, 2016

< 1% match (publications)
Arlie O. Petters, Xiaoying Dong. "An Introduction to Mathematical Finance with Applications", Springer Science and Business Media LLC, 2016

< 1% match (student papers from 05-Mar-2007)
Submitted to City University of Hong Kong on 2007-03-05

< 1% match (student papers from 21-Apr-2010)
Submitted to City University of Hong Kong on 2010-04-21

< 1% match (student papers from 03-Aug-2018)
Submitted to City University on 2018-08-03

< 1% match (student papers from 10-Jun-2011)
Submitted to Associatie K.U.Leuven on 2011-06-10

< 1% match (student papers from 12-Nov-2019)
Submitted to Kabarak University on 2019-11-12

< 1% match (student papers from 09-Aug-2019)
Submitted to Coventry University on 2019-08-09

< 1% match (student papers from 17-Oct-2019)
Submitted to All Hallows High School on 2019-10-17

< 1% match (student papers from 28-Aug-2012)
Submitted to Leeds Metropolitan University on 2012-08-28

< 1% match (student papers from 10-May-2013)
Submitted to University of Warwick on 2013-05-10

< 1% match (student papers from 25-Oct-2017)
Submitted to Erasmus University of Rotterdam on 2017-10-25

< 1% match (student papers from 29-Oct-2019)
Submitted to Erasmus University of Rotterdam on 2019-10-29

< 1% match (student papers from 20-Dec-2017)
Submitted to An-Najah National University on 2017-12-20

< 1% match (student papers from 20-Apr-2010)
Submitted to City University of Hong Kong on 2010-04-20

< 1% match (student papers from 22-Jul-2016)
Submitted to Universitat Mannheim on 2016-07-22

< 1% match (student papers from 11-Dec-2008)
Submitted to University of Wales, Bangor on 2008-12-11

< 1% match (student papers from 05-Nov-2010)
Submitted to University of Southern California on 2010-11-05

< 1% match (student papers from 28-Aug-2012)
Submitted to University of Edinburgh on 2012-08-28

Department of Electrical Engineering FINAL YEAR PROJECT REPORT BENGEGU4 -CDE- 2019/20 -CWT- 01 Project Title: Portfolio Management with Heuristic Optimization Student Name: YOGISH Suhas Student ID: 54482837 Supervisor: Mr TING, Van C W Assessor: Mr NG, K T Bachelor of Engineering (Honours) in Computer and Data Engineering Student Final Year Project Declaration I have read the student handbook and I understand the meaning of academic dishonesty, in particular plagiarism and collusion. I declare that the work submitted for the final year project does not involve academic dishonesty. I give permission for my final year project work to be electronically scanned and if found to involve academic dishonesty, I am aware of the consequences as stated in the Student Handbook. Project Title : Portfolio Management with Heuristic Optimization Student Name: YOGISH Suhas Student ID: 54482837 Signature: Suhas Yogish Date: 20/04/2020 No part of this report may be reproduced, stored in a retrieval system, or transcribed in any form or by any means – electronic, mechanical, photocopying, recording or otherwise – without the prior written permission of City University of Hong Kong. Turnitin Originality Report: Abstract Portfolio Management is the art of choosing and managing a group of financial securities such as bonds or equity instruments, derivative instruments like futures and forward contracts etc. There many key elements to portfolio management, such as asset allocation, diversification of the portfolio, and rebalancing the portfolio. In this project, my objective is to develop strategies to build optimal portfolios which aim to outperform index funds such as the S&P500 Index by researching and developing different heuristic methods for the key elements of portfolio management. One of the strategies use various fundamental factor information such as return on an asset, asset turnover, net margin etc., to build alpha factors. The alpha factors are used as input variables in a machine learning clustering model which is used to diversify and build portfolios from the stock universe. Another strategy uses Piotroski F-scores which is calculated using financial metrics which can be grouped into three criteria's, leverage, profitability and liquidity to filter out companies with bad fundamentals. This is combined with machine learning models which predict the future movements of price and volume for the value stocks which is used to decide whether to trade long or short. Different optimization techniques such as mean-variance optimization and minimax optimization are studied and tested on the trading strategies to allocate optimal weights to the underlying assets in the portfolios. These strategies are developed using python programming language and using the Quantopian API for data collection and back-testing in a trading environment. i Acknowledgements I would like to express my deep sense of gratitude and thanks to my final year project supervisor, Mr. Van Ting for his constant guidance, encouragement and feedback throughout the two semesters. His guidance and recommendations played a significant role in the success of this project. His detailed feedbacks helped me to achieve better results and improve the overall quality and performance of this project. Next, I would like to thank my mentors and managers during my various internships who introduced me to the field of data science and finance. It has been a great learning experience so far and I will continue this journey into my future career. I would also like to express my thanks to the EE department at City University of Hong Kong for providing a world class educational experience the past four years of my undergraduate studies and all the professors who have taught me during this time. Finally, I would like to thank my parents and family for giving me the opportunity to study in Hong Kong and for the constant motivation to achieve more every day and work hard along the way. Their support has been instrumental in shaping my adulthood and making me who I am today. To say the least, I'm excited and looking forward to opening a new chapter in my life after graduation. ii 1. Table of Contents

 1.
Introduction In the world of finance, portfolio management and optimization are a problem which is solved everyday by asset managers in the field. Portfolio management and optimization is the process of allocating capital across different financial assets such as equities, bonds and mutual funds to build investment solutions. While building a portfolio, the investor must have an investment objective and must address various factors such as how much risk they are prepared to take, the duration of their investment and many other factors which will be discussed in this paper. Over the past decade, there has a been exponential growth and research in the field of artificial intelligence and machine learning. Advanced algorithms are being developed and tested in multiple areas and finance is one of the front runners. In this project, I will tackle portfolio optimization with help of different machine learning models and statistical models to optimally allocate weights to various equity assets and build portfolios which beat an index benchmark. The algorithms will be back tested in a real-world trading environment using Quantopian, an online quantitative research tool which provides back testing framework and also numerous APIs to get financial data. Building portfolios with the goal of beating an index benchmark is commonly known as Active Portfolio Management in the industry. Various types of fundamental analysis and techniques for market research and company stock analysis are discussed and implemented in the strategies built. Market factors and style factors are used in the alpha discovery and research stage to build an alpha which has good predictability. Quantopian, the online quantitative analytics and backtesting platform is used to conduct research on historical datasets, alpha research and more importantly to build the trading strategies and backtesting for evaluating performance. 1 2. Background 2.1. Fundamentals of Investment 2.1.1. Asset An asset is essentially anything which possesses monetary value. Some example of assets include cash, checking and savings accounts, treasury bills, stocks, bonds etc. In this project I will be focusing on stocks. 2.1.2. Stocks A stock or equity is an asset or a form of security which indicates a proportionate ownership in a company by the holder of the stock. Stocks are generally issued by corporations to raise funds in order to operate their businesses. Stocks are bought mostly on stock exchanges. Some common and famous stock exchanges are the New York Stock Exchange (NYSE), the London Stock Exchange (LSE), NASDAQ to name a few. Stocks are usually the foundation of almost

every portfolio built. 2.1.3. Portfolio A portfolio is a set or group of publicly traded assets such as equities, bonds, mutual funds, commodities etc. Sometimes portfolios can also include non-publicly traded assets such as real estate and art. Investors construct their portfolios based on their risk tolerance their investment objectives and these portfolios are generally managed by the investors themselves or wealth managers. 2.1.4. Portfolio Construction and Optimization Portfolio Construction is all about selecting the right assets based on the investor's objectives and risk tolerance. The most important aspect of constructing a portfolio is the asset allocation. Breaking down the portfolio optimization process, there are four important steps one must consider: • Risk Profiling - High Risk or Low Risk based on a scale determined by the investor. • Asset Allocation - The right combination of assets for the portfolio. • Fine-Tuning the Portfolio - Based on the risk profile and the existing assets, add or remove the underlying assets most suitable to the investor, potentially reducing the risk and increasing the returns. 2 • Rebalancing - The investor must review the portfolio from time to time and rebalance the weight allocated to the underlying asset based on the market. 2.2. Fundamentals of Portfolio Theories 2.2.1. Modern Portfolio Theory In 1952 Harry Markowitz published a paper called "Portfolio Selection" where he introduced the Modern Portfolio Theory (MPT). MPT theorizes on how a risk-averse investor can build and optimize their portfolios to achieve maximizing expected returns for a given level of market risk. 2.2.2. Portfolio Expected Returns In a portfolio of N assets, the expected returns are calculated as the weighted average of the individual returns of all N assets. For example, if a portfolio contains 5 equally weighted assets, each with individual returns of 6%, 7%, 12%, 4% and 15% respectively, then the expected returns of the portfolio will be calculated by: E(R) = (0.06 x 0.20) + (0.07 x 0.20) + (0.12 x 0.2) + (0.04 x 0.20) + (0.15 x 0.20) = 8.8% 2.2.3. Efficient Frontier Every possible weighted combination of assets in a portfolio can plotted onto a graph showcasing the portfolio risk on the X-axis and the portfolio expected return on the Y-axis. Based on this graph an investor can choose the most desirable combination of assets. The upward sloping hyperbola which connects all the most efficient portfolio asset combinations is called the efficient frontier. For an investor following the Modern Portfolio Theory, investing in a portfolio not on the efficient frontier is not desirable. 2.2.4. Sharpe Ratio Nobel laureate William Sharpe introduced a new method to understand a portfolio's performance. The Sharpe Ratio is used by investors to understand the return on investment (ROI) compared to the amount of risk undertaken. The Sharpe ratio can be calculated by using the formula: 3 $SharpeRatio$ = !!"!" #! Where: Rp = $ExpectedReturnofPortfolio$ $Rf = RiskFreeRate$ $\sigma p = Standard\ Deviation\ of\ Portfolio's\ Excess\ Returns$ 2.3. Active Portfolio Management In the industry there two main types of investment management strategies practiced by portfolio managers, passive portfolio management and active portfolio management. Passive investing generally involves replicating the security holdings of an index fund to achieve similar results. Active investing on the other hand involves managing portfolios in order to beat an index or a benchmark to achieve superior returns. As the name suggests, in active investing a portfolio manager trades more frequently and conducts more in-depth fundamental research. Active portfolio management does involve high risk compared passive. In my project my goal is to implement active portfolio management and build strategies aimed to beat the Standard & Poor's 500 index [1]. 2.4. Fundamental Factor Research [2] Financial factors are essentially the foundations of active investing. They refer to various ratios and metrics which are used to measure a company's financial characteristics which are derived from financial statements, quarterly reports, balance sheets, etc. Some examples of such factors are market capitalization, net income, net margin and cash flow. I will make use of similar fundamental factors in one of strategies that help to determine characteristics that affect the portfolio's assets risk and return. One of the important steps is to select the significant factors which can be done by conducting in-depth factor research [2] [3] [4]. 2.5. Factor Based Strategies With respect to investing strategies, there are two main types of factors that have driven returns, macroeconomic factors and style factors. Some common style factors used in portfolios include: 4 a. Momentum: This factor identifies stocks which have performed well in the past and tends to exhibit high returns in the future. b. Value: This factor identifies stocks which have low prices when compared to their fundamental factors. This is used in determining different metrics like the Piotroski score explained ahead. c. Volatility: This factor is essentially the standard deviation or risk of a particular asset. Various research studies have shown that stocks with lower volatilities earn higher risk adjusted returns. d. Size: This is one of the most important factors which can used to diversify a portfolio. The size of an asset is captured by the market capitalization. Historically, small -cap stocks exhibit higher returns compared to large- cap stocks. These factors are used to help improve portfolio outcomes, reduce the portfolio volatilities and diversify the portfolio assets [4]. I make use of style factors in my algorithm as input variables for my machine learning model. 2.6. Piotroski Score The Piotroski F-Score is a score used to determine the strength of company based on their financial statements. This score is used to filter out stocks of weak companies and determine the top value companies. The score is awarded from zero to nine based on a few criterions where zero is the worst and nine is the best. The criterions used to calculate the score are as follows [5]: Profitability (4 Points): • Positive net income. • Positive return on the underlying asset. • Positive cash flow. • Cash flow being greater than net income. Leverage and Liquidity (3 Points): • Decreased leverage compared to previous year. • More liquidity compared to previous year. • No new shares issued in the previous year. 5 Operational Efficiency (Two Points): • Increased gross margin compared to previous year. • Increased asset turnover compared to previous year. Based on the above criterions, a stock universe can be ranked and the top ranking stocks can be selected for further diversification or constructing a portfolio [5]. This method will be further explained in the methodology and implementation section of the report. 2.7. Machine Learning Models 2 .7.1 K-Means Clustering Clustering analysis is an unsupervised machine learning method which is used for finding homogenous groups from the data in a way that the objects in each group share similar characteristics to each other when compared to other group objects. There are many mathematical clustering algorithms designed which have different definitions of how clusters are calculated such similarity or distance functions like Euclidean distance or correlation distance and how many number of clusters are created. Clustering algorithms can be divided into two main types of algorithms: partitional clustering and hierarchical clustering. For my project I will be implementing partitional clustering algorithms which creates partitions in the data by grouping disjoint clusters since there is no hierarchical nature to stock data. K-Means Clustering algorithm is a commonly used partitional algorithm where it groups n observations into k number of clusters where each observation belongs to the cluster centroid with the nearest mean. The objective function of the algorithm below is to be minimized which is the cluster sum of squares: * *arg*$ min 0 0‖*x* − *μ*%‖& %+, '∈)# Where x are the number of observations in the data, S = S1 , S2 ,...., Sk are the set of observations and μi is the mean of each points in Si 6 K-Means is an iterative algorithm which selects the centroids of each group randomly in the first iteration which are used as the initial points for each cluster. From here it optimises the positions of the centroid with repetitive calculations. The approach for calculating the positions in every iteration is called Expectation-Maximization and can be mathematically defined by the equation above. The figure 1 below shows clustering example of stock returns data where we have three clusters and each cluster with a centroid. Figure 1 - KMeans Clustering example with centroids The algorithm is complete when the number of iterations are completed or when the centroids have no change in their values, meaning they are stabilised. This entire process can be shown graphically by the below diagram figure 2. Figure 2 - KMeans Clustering Flowchart https://www.researchgate.net/figure/K-Means-Clustering-Flow- Chart_fig1_269800441 7 2.7.2. Support Vector Machines Support vector machines is a machine learning model which is used for classification and regression tasks but more widely used for classification. The objective function of a support vector machine is to calculate the hyperplane in an N-dimensional space where N is the number of input features. The hyperplane is calculated in an iterative manner in order to minimize the error [6]. The three main concepts in an SVM algorithm are: • Margin - This is the gap between the closest datapoints of two separate classes. The objective of the algorithm is to maximise the margin by finding the largest distance. • Support Vectors - Support Vectors are datapoints that are closer to the hyperplane and influence the position of the hyperplane. They are used to maximise the margin of the classifier. Figure 3 - SVM Margin and Vectors • Hyperplanes - They are the decision boundaries which help to distinguish between datapoints and classify them accurately. Dimensions of hyperplanes are dependent on the number of input features used to train the SVM. 8 Figure 4 - SVM Hyperplanes in 2D and 3D 2.7 .3. Gradient Boosting Regressors Gradient Boosting algorithm is another type of machine learning technique which is used for both classification and regression problems. Gradient Boosting Regressors (GBR) is a form of ensemble tree model, generally decision trees. The reasons for difference between the actual and predicted values when using a machine learning technique is mainly due to noise, variance and bias. With the help of ensemble models like GBR, these factors can be reduced [7]. Boosting is an ensemble technique where the predictors are made sequentially where subsequent predictors learn from the mistakes of the previous predictors. This helps to reduce the bias and variance. The intuition behind the GBR algorithm is to repetitively leverage the patterns in residuals and strengthen the model with weak predictions and make it better [7]. The objective of the GBR algorithm is to minimize the defined loss function. Loss functions are defined based on the problem being solved and for a regression task the loss function can be the mean squared error [7]. 9 2.8. Portfolio Optimization Techniques 2.8.1. Mean Variance Optimization Mean-Variance optimization technique was first introduced in Harry Markowitz' s Nobel Prize winning work on 'Modern Portfolio Theory'. An investor who follows the MPT for their portfolio, models the rate of return on the underlying assets as a random variable [8]. The objective of the optimization is to choose the weights for the assets in the most optimal method. Mean Variance optimization uses standard deviation of a portfolio as the measure of risk. The inputs for the optimization model are as follows: • The expected returns for each asset in the portfolio. • The standard deviation (risk) of each asset in the portfolio. • The correlation matrix between each asset in the portfolio. The output of the optimizer is called the Efficient Frontier. The efficient frontier can be visualized on a graph which shows a set of optimal portfolios from which an investor can choose the portfolio which offers the highest expected returns for a given level of risk or the lowest risk for a given level of returns [8]. The optimization algorithm for a portfolio of two securities j and k with data points {yjt, ykt} can be mathematically denoted as: $ $ *min*! $ $ *w*"*w*#*s*"# "%& #%& Parameter Constraints: $ $*w*"*y*(" ≥*G* "%& Where G is the minimum level of expected returns of the portfolio. 10 $ $ *w*"< *W* "%& Where W is the maximum weight allocation available which is 1 (100%). *S*"# = ()$ ∑(*%&(*y*"* − *y*(")(*y*#* − *y*(#) & Where Sjk is the covariance matrix between the two securities. 2.8.2. Mini-Max Optimization Mini-Max optimization is a simple linear programming algorithm which chooses the optimal portfolio based on the minimum return rather than minimum variance like the mean-variance optimization technique. Being a linear programming algorithm, Mini-Max makes it possible to constrain some variables as Boolean 1-0 or integer values. Other than the computational convenience, it also has logical advantages when the returns of the assets are non-normally distributed [8]. Suppose we have a portfolio with N assets over a time T, let: *y*-. = *Return on asset j at a particular time t y*>- = *Mean of returns of asset j w*- = *Weight allocated to asset j* 0 *y*/. = 0 *w*-*y*-. = *Return on portfolio at time t* -+, 0 *E*/ = 0 *w*-*y*>- = *Average Return on portfolio* -+, 11 *M*/ = *Minimum Return on portfolio* The optimization maximizes the minimum return Mp , that is it minimizes the maximum loss where loss is the negative gain, in other words maximize the minimum gain. This can be derived as follows: *max M/ Subject to the following constraints*: 0 0 *w-y-.* − *M/* ≥ 0, *t* = 1, 2, ... , *T* -+, 0 0 *w-y>-* ≥ *G where G is the minimum expected returns* -+, 0 0 *w-* ≤ *W where W is the maximum weight allocation* -+, 2.9. Quantopian Environment Quantopian is an open source quantitative research and analytics tool which provides numerous APIs and a back testing environment for trading algorithms. For my project I will be using Quantopian mainly for alpha research and back testing since it makes it easier to access good quality financial data through their API and also they provide more computational power than compared to my local machine since it is hosted online. 3. Objective For my final year project, my objective is to learn the fundamentals of quantitative finance and machine learning and apply my knowledge into building trading strategies which can beat the benchmark index which is the Standard & Poor's 500 index. Over the course of the project I built and tested various strategies trying to beat the benchmark. I have finally chosen two strategies to present which not only beat the benchmark but also achieved a relatively high Sharpe ratio. 12 4. Literature Review and Related Work 4.1. Portfolio Diversification using Cluster Analysis A lot of the research work previously achieved in this field has received positive sentiments from the industry experts. Correlation is a fundamental statistical method which gives us the likelihood that two variables move together. A positive correlated pair essentially means when one goes up the other goes up too or vice versa. A negative correlated pair is the opposite, when one moves up the other moves down. Applying this to the stock market, investors try to build portfolios with negative correlated assets so that they can diversify their portfolios and are less subjected to market risks. Correlation based clustering of the Stockholm Stock Exchange by Frederick Rosen (2006) is good example of a research paper where the author classifies stocks based only on the correlation between them. In this method of clustering if one clusters stock price decreases, another clusters stock price will not decrease thereby hedging to reduce loss. Correlations between stocks is not the most efficient way to diversify a portfolio since they often reverse or change during periods of economic stress. Therefore during the times of economic stress the assets within a cluster may not be highly correlated with each other or clusters may not be negatively correlated with each other. This is one drawback from clustering based on correlations and defeats the purpose of diversification from an investor standpoint by exposing the portfolios to higher risk [9]. 4.2. Piotroski F-Score to identify value stocks As I mentioned in the background study, Piotroski F-Score is used to identify high value stocks and reduce an investors stock universe. There have been various researchers who have developed and tested numerous methodologies to improve the returns of a portfolio in the value universe. Joseph Piotroski (2000) tested his investment strategy between the time periods 1976 and 1996 using American stocks. Piotroski calculated the book-to-market at the end of the financial year for each firm and assign them to a quintile. He then calculated the F-score for each firm in the highest quintile assigning them a score between 0-9 where 9 is the highest value firm. He created two portfolios, one containing low F-score stocks and other containing high F-score stocks. He documented that the high F-score portfolio outperformed the low F- score portfolio [10]. One of the most useful benefits of using this method devised by Piotroski is that this strategy helps in identifying small and mid-capitalization companies which have 13 high value. These companies generally have low fundamental analysis coverage and neglected by investors [11]. 4.3. Minimax and Mean Variance Optimization There are many portfolio optimization techniques used in the industry. Markowitz (1952) introduced the mean-variance optimization technique in his Nobel prize winning essay on Modern Portfolio Theory. A Minimax Portfolio Selection Rule with Linear Programming by Martin Young (1996) introduces a new optimization technique which uses historical returns data similar to mean-variance and is a solution to a linear programming problem. This optimization technique uses the minimum returns of an asset as the measure portfolio volatility as compared to the mean-variance optimization which uses the variance [8]. 5. Strategy Design and Implementation 5.1. K-Means Clustering with MiniMax Optimization (Strategy 1) This is the first trading strategy that I developed in which I implement K-Means clustering for portfolio diversification and use the Mini-Max optimization algorithm for optimal weight allocation. 5.1.1. Alpha Research In this strategy I use four different factors equi-weighted into one alpha factor which is used as the input feature for the K-Mean clustering algorithm. Choosing the right factors requires a lot of research to understand how the factors behave and affect a portfolio. Quantopian has a research environment where we can use a tool called 'AlphaLens' for analysing a factor's effectiveness at predicting future returns. I can also create visualizations which help to further understand a factor. For my algorithm I researched and decided to the following four factors as my input feature: a. Return on an asset = Net income / Total assets b. Asset Turnover = Net Revenue / Total assets c. Net Margin d. Growth Score 14 Once the factors are decided, I combine them together with equal weights to form one alpha in order to analyse and check if our factor data affects prices in the future. The time periods checked are 1 day, 5 days and 10 days, denoted as 1D, 5D and 10D respectively in the visualizations. The stock universe is split in quantiles where quantile 1 contains the 20% of the stocks with the lowest alpha factor values and quantile 5 with the highest 20%. I then research the returns generated by each quantile over a time period. The goal is to make sure that quantile 5 performs the best while quantile 1 is the worst. Figure 5 - Mean Period Wise Returns for Factor Quantiles Figure 5 shows the mean returns for each quantile for the time period mentioned before. We can see that the first quantile performance is negative while the fifth quantile has a positive mean return. This is one of the many analysis we can conduct to see whether our alpha factors are good for our portfolios. Figures 6, 7 and 8 plot the log cumulative returns of the each of the quantiles over a five year back testing period from 2015 to 2019. The visualization shows that the fifth quantile's total cumulative returns is much higher and outperforms the first quantile. 15 Figure 6 - Cumulative Factor Returns for 1D Forward Returns Figure 7 - Cumulative Factor Returns for 5D Forward Returns Figure 8 -

Cumulative Factor Returns for 10D Forward Returns 16 Another important metric to determine our alpha factor performance is the information coefficient (IC). The IC is used to describe the correlation between the predicted and actual returns and sometimes also evaluate the skill of a portfolio manager or quantitative investment analyst. The score is between -1 to 1 and quantifies the predictiveness of the alpha factor. Figure 8 shows the IC decay for the alpha factor over the period of one trading year (252 days). The alpha factor is considered useless when the IC drops below 0. However, in my alpha factor, we can see the IC score gradually increase over time. Figure 9 - Information Coefficient Decay for Alpha Factor 5.1.2. Data Collection The factor data, stock universe data, stock prices data are all collected from a Quantopian API. This makes it data collection faster and much simpler in terms of data pre-processing and cleaning. After the alpha research phase, I started the data collection for the factors and building the foundation for my trading strategy. The stock universe used for my strategy is the Quantopian 1500 stock universe which provides a list of the top 1500 tradeable equities where tradeable can be defined as: Stock must be the primary share class for its company. • The company must have a known market capitalization. • The stock must not be a depository receipt. • The stock must not be traded over the counter. 17 • The stock must have a previous closing day price. • The stock must have a volume greater than zero on the previous trading day. With the stock universe available to choose from, I add another parameter to filter only the companies issuing stocks with a market capitalization of greater than $1.5 billion USD which is the median market cap for the S&P 600 index so that I can include small, mid and large market cap companies in my universe. 5.1.3. Clustering Algorithm Once the factor data is collected, the clustering function creates an equi-weighted alpha factor similar to what was done in the research stage. This alpha factor is used as the input feature for the KMeans clustering function. The KMeans function has the hyperparameters set to n_clusters = 15, n_jobs = 1 and random_state = 10. The algorithm outputs a cluster for each asset in the stock universe and all single stock clusters are removed if there are any. Once the clustering algorithm is complete, high performing stocks are filtered from each cluster by calculating the Sharpe ratio of all the stocks by taking the rolling price data of three-month period. The highest Sharpe ratio stock from each cluster are selected for the portfolio. 5.1.4. Weight Optimization Once the stock universe is selected, I calculate the optimal weights for each asset in the portfolio by using the Mini-Max optimization technique previously discussed. The optimizer has maximum weight of 5% per asset as one of the constraints set since I don't want to have a high exposure for any single asset. The optimization algorithm uses the historical daily returns for a two-month period as one of the inputs and outputs the target weights for each asset. 5.1.5. Portfolio Rebalancing and Trading The portfolio is rebalanced every month by calling the clustering and weight optimization functions. Depending on the target weights, an asset is traded long or short. The initial investment for trading is $1,000,000 USD. The rebalancing takes place three minutes after the market opens on the first day of each month. 18 5.2. Piotroski F-Score value stocks (Strategy 2) This is the second strategy that I backtested which beat the benchmark index. In this strategy I filter high value companies using the Piotroski f-score method and then predict whether to go long or short using a machine learning algorithm and historical data. The weights are optimized using mean-variance optimization. 5.2.1. Data Collection Similar to the previous algorithm, the stock universe used is the Quantopian 1500 stocks list. To improve the performance of the algorithm, I conduct a few stock valuations analyses on the stock universe and exclude stocks which don't meet the below selection criteria's: a. Companies with market capitalization greater than $1Billion USD b. Companies with positive earnings before interest and tax (EBIT). This metric indicates a company's profitability and is generally used to analyze the performance of a company's operations [12]. c. Companies with positive enterprise value (EV). This metric is a measure of a company's total value where EV = Market Cap + Total Debt - Cash (liquid assets). Enterprise value is used as a foundation for many financial ratios used to measure performance [13]. d. Companies with positive EV/EBITDA ratio. This metric compares a company's enterprise value to the earnings before tax, interest, depreciation and amortization. This ratio helps in identifying undervalued companies [14]. e. Companies with low annualized volatility stocks. Selecting stocks with low volatility helps in reducing the beta of the overall strategy performance. Once the universe is defined, the Piotroski f-score is calculated for each stock in the universe. The pipeline returns the stocks and their respective scores. 19 5.2.2. Machine Learning Model The top fifty stocks with the highest Piotroski score are selected for the portfolio rebalancing every month. Two machine learning models are trained using the historical data of each stock. The first model is used to train the historical pricing data of a stock so that the present data can be used to predict whether a stock price will be increasing or decreasing. This information can be used to decide whether to go long or short the stock. Similarly, the second model uses the historical volume data of stocks to predict the change in volume. The historic range used as training data is set to 180 days (6 months). For the prediction data, the lookback period is 2 days since the latest pricing and volume data is important to predict the future movements. The machine learning model used is the Gradient Boosting Regression algorithm discussed before. 5.2.3. Weight Optimization Once the machine learning model predicts the price movement of the stock and whether to go long or short, the optimal weights are calculated using mean-variance optimization previously discussed in detail. The optimization algorithm uses the historic two-month pricing data to calculate the optimal weights for the portfolio. 5.2.4. Portfolio Rebalancing and Trading The rebalancing period is set to the beginning of each month, three minutes after the market opens similar to the previous strategy. The initial investment used for the algorithm is $1,000,000 USD. If both the machine learning models predictions are greater than zero then the stock is traded long with the optimal weight, if both the predictions are lesser than zero then the stock is traded short and finally if one prediction is greater than zero and the other lesser, then the stock is not traded or rebalanced for that month. 20 6. Results and Performance 6.1. K-Means Clustering with MiniMax Optimization (Strategy 1) The strategy was back tested for a period of 5 years from 2015-01-01 up until 2019-12-31. From Table 1 below, the overall performance of the strategy during the backtest period can be deduced. Table 1 - Strategy Performance metrics (Strategy 1) Performance Metrics Backtest Annual Returns Cumulative Returns Annual Volatility Sharpe Ratio Sortino Ratio Maximum Drawdown Daily Turnover Alpha Beta 16.828% 117.233% 12.574% 1.30 2.00 -13.645% 7.962% 0.12 0.39 The cumulative returns of the strategy are 117.233% compared to the benchmark S&P500 Index returns of 72.82% during the time period. The strategy also has an alpha score of 0.12 or 12% which is the excess return compared to the benchmark and risk undertaken. The overall Sharpe Ratio is 1.30 which is considered to be good by investors. 21 It’s also good to understand how the strategy is performing with respect to factors such as size, momentum, value and also for each sector. This is useful so that the strategies can be optimized further to get better results. Table 2 - Performance related to common risk (Strategy 1) Table 2 - Exposure Summary (Strategy 1) 22 Based on the cumulative returns performance for the 11 sectors and the 5 style factors seen on table 2, I plot the returns of the factors and sectors on a graph as seen in the figures 10 and 11 below. Momentum and volatility are the best performing style factors while sectors such as consumer cyclical which includes industries like automobile, housing, entertainment and retail, financial services sector and technology sector stocks generate superior returns. Figure 10 - Cumulative returns for sectors (Strategy 1) Figure 11 - Cumulative Returns for Style Factors (Strategy 1) Figure 12 below plots the rolling Sharpe ratio over the period of the backtest. From the graph one can understand the performance of the algorithm at various points in time. Analysing the plot, we see that the strategy performs very well during 2016 with the Sharpe ratio peaking, however during early 2019 all the way to late 2019 we see a drop in the Sharpe ratio. This may be due to exposure to various risk factors which can be further analysed to improve the algorithm. 23 Figure 12 - Rolling Sharpe Ratio (Strategy 1) Based on the understanding of the strategies Sharpe ratio over the time period, the annual returns graph per year in figure 14 validates the assumption made above. The highest annual returns are recorded in the year 2016 and lowest in the year 2018. Figure 13 plots the cumulative total returns along with the common returns and the specific returns which is the difference between the total returns and the common returns. Figure 13 - Total, Specific and Common Returns (Strategy 1) 24 Figure 14 - Annual Returns per year (Strategy 1) Figure 13 - Strategy Performance against Benchmark (Strategy 1) 25 Figure 14 - Monthly performance Heat Map (Strategy 1) Figure 13 and 14 are visualization illustrating the strategy’s returns which is helpful in gauging how the algorithm performs in different time periods throughout the backtest. It also helps in identifying seasonal patterns if any and optimize the strategy. In figure 14, the top performing month of the strategy can be identified as July 2015 with 20% returns. The drawdown of the strategy is calculated to be -13.6% over the 5 year backtest period. This means the peak to trough decline of the strategy was 13.6%. To put this into perspective, during the 2008 financial crisis, the market drawdown was over 50%. Drawdown is a good metric to measure the financial risk of a strategy. According to Quantopian a good benchmark for maximum drawdown is less than 20%. Figure 14 shows the top 10 drawdown periods for the strategy how it performed during the drawdowns and Figure 16 is an underwater plot to get a better visualization into the drawdowns. Table 3 gives more information about the drawdowns like the duration in number of days and the net drawdown percentage for each period. 26 Table 3 - Top 5 worst drawdowns (Strategy 1) Figure 15 - Top 10 drawdowns (Strategy 1) Figure 16 - Underwater Drawdown (Strategy 1) 27 6.2. Piotroski F-Score value stocks with Machine Learning (Strategy 2) This strategy is backtested similarly to strategy one on the Quantopian environment for a four- year period from 2015-01-01 to 2018-12-31. The backtest performance for the strategy can be found the table below with a comparison with strategy 1. Table 4 - Strategy 2 Performance Metrics Performance Metrics Backtest Annual Returns Cumulative Returns Annual Volatility Sharpe Ratio Sortino Ratio Maximum Drawdown Daily Turnover Alpha Beta 31.853% 201.254% 20.291% 1.46 2.40 -16.554% 2.245% 0.28 0.24 Looking at the comparison column, the strategy 2 annual returns is almost twice as strategy 1 with little higher annual volatility. This explains the higher Sharpe ratio as well. Strategy 2 experiences a similar maximum drawdown, a higher alpha score and a lower beta score which are all signs of a better performance. 28 Table 4 and 5 is the performance relative to common risk factors and also exposures to style factors and sector information. Figure 16 plots the cumulative returns, specific returns and the common returns. Table 5 - Performance related to common risk (Strategy 2) Table 6 - Exposures to Sector and Style Factors and Common Returns (Strategy 2) 29 Table 6 shows that the technology sector and the financial services sector generate high positive returns and the volatility style factor has the highest returns out of the five factors while consumer cyclical sector stocks and short-term reversal factor have negative cumulative returns. Figure 17 - Cumulative Returns of Style Factors (Strategy 2) Figure 18 - Cumulative Returns of Sectors (Strategy 2) 30 Figure 19 - Cumulative, Specific and Common Returns (Strategy 2) From figure 19 we can understand that the strategy performs well with respective common risk factors which are the 11 sector and 5 style factors shown in table 5 above. These risk factors make up the common returns. The specific returns are the difference between the total returns and common returns. Figure 20- Cumulative Returns Against Benchmark (Strategy 2) 31 Figure 21 - Heatmap of monthly and annual returns (Strategy 2) Figure 21 shows the best and worst performing months for the strategy (green to red). January 2017 and July 2016 were top performing months with 17% and 16% monthly returns, while November 2016 and November 2015 were the worst with -6.4% and -6.3% returns respectively. Figure 22 - Annual Returns Histogram (Strategy 2) 32

Figure 23 - Rolling Sharpe Ratio (Strategy 2) The Sharpe ratio sees a sharp decline starting from the end of 2017 to the middle of 2018 in figure 23. This period can be further analyzed looking at the maximum drawdown visualizations in figure 24 and 25. Figure 24 plots the top 10 worst drawdown periods along with the cumulative returns during that period. Figure 24 - Top 10 Drawdown periods (Strategy 2) 33 Figure 25 - Underwater Drawdown (Strategy 2) Figure 25 is an underwater plot to help quantify the drawdowns during the backtesting time period. We can see the longest and highest maximum drawdown occurred in 2018. To get more information about the drawdowns, table 7 contains information about the net drawdown percentage, the duration in number of days, the peak, trough and recovery dates. Table 7 - Top 5 worst drawdowns (Strategy 2) 34 6.3. Comparison of Strategies Although both the strategies beat the S&P500 benchmark index, they have different results in terms of performance, cumulative total and specific returns, exposures to sectors and factors, volatilities, etc. Table 8 is a comparison of both the strategies performances during the time period 2015-01-01 to 2018-12-31. Figure 26 plots the cumulative returns of both the strategies. Table 8 - Comparison of Strategy Performances Performance Metric Strategy 1 Strategy 2 Cumulative Returns 102.12% 201.25% Annual Returns 19.30% 30.85% Annual Volatility 13.31% 20.30% Sharpe Ratio 1.39 1.46 Maximum Drawdown -11.78% -16.55% Alpha 0.15 0.28 Beta 0.40 0.24 Looking at table 8, the strategies have different risk-returns ratios. Strategy 1 would seem more suitable for an investor who is looking for an active investment strategy but willing to take same level of risk as compared to the S&P500 which has annualized 5Y returns of 6.6% with a standard deviation of 13.50% [15]. Strategy 2 is suitable for an investor with a higher appetite for risk. Annual returns of 30.85% which is almost 5x higher than the S&P500 annual returns for a slightly higher risk appetite in comparison. 35 Figure 26 - Cumulative Returns of Strategy 1 & 2 7. Conclusion and Further Work The project was successfully completed with both the trading strategies meeting the objectives of beating the S&P 500 benchmark index by implementing different heuristic mathematical optimization techniques such the minimax optimization and the mean-variance optimization along with machine learning models used to diversify portfolios, predict the future change in price and volume. The results of the strategies show promise of utilizing machine learning techniques in portfolio management. The strategies can be further optimized with extensive fundamental research to build better alpha factors, reduce gross exposures to different underperforming sectors and style factors in order to achieve higher risk-adjusted returns. The strategies can be tested over longer time periods and also during financial crisis periods to evaluate the performances when the market is different. I will be further developing the algorithms and strategies for the Quantopian trading algorithm contest. 36 Although there is no widely used machine learning strategies for active portfolio management, I hope this project and research study can be utilized in the practice as a beginning step and be further developed and improved. Further research can be conducted on different machine learning approaches and hyperparameter optimization for better results. The strategies are evaluated using historical data and backtesting to verify how the strategy would have performed. This is only the initial stage of the evaluation process to research and ensure the strategies are fundamentally strong and sound. The strategies need further enhancements before an investor uses them in a live trading environment. 37 Appendix I: 38 Bibliography

[1] N. Lioudis, "Passive vs. Active Portfolio Management: What's the Difference?," 15 February 2020. [Online]. Available: https://www.investopedia.com/ask/answers/040315/what-difference-between-passive- and-active-portfolio-management.asp. [2] T. Segal, "Fundamental Analysis," 16 March 2020. [Online]. Available: https://www.investopedia.com/terms/ f/fundamentalanalysis .asp. [3] "What is factor investing?," BlackRock, [Online]. Available: https://www.blackrock.com/us/individual/investment-ideas/what-is-factor-investing. [4] J. Chen, "Factor Investing," 16 March 2020. [Online]. Available: https://www.investopedia.com/terms/f/factor-investing.asp. [5] J. Chen, "Piotroski Score," 22 June 2019. [Online]. Available: https://www.investopedia.com/terms/p/ piotroski-score .asp. [6] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," 08 June 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine- learning-algorithms-934a444fca47. [7] P. Grover, "Gradient Boosting from scratch," 09 December 2017. [Online]. Available: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d. [8] M. Young, "A Minimax Selection Portfolio Rule with Linear Programming Solution," Ann Arbor, 1996. [9] K. Marvin, "Creating Diversified Portfolios Using Cluster Analysis," 2015. [10] S. Bülow, "The Effectiveness of Fundamental Analysis on Value Stocks – an Analysis of Piotroski's F-score," 2017. [11] J.-H. M. Mohr, "Utility of Piotroski F-Score for predicting Growth-Stock Returns," Frankfurt. [12] C. B. Murphy, "Earnings Before Interest and Taxes – EBIT," 02 April 2020. [Online]. Available: https://www.investopedia.com/terms/e/ebit.asp. [13] C. B. Murphy, "Enterprise Value – EV," 20 June 2019. [Online]. Available: https ://www.investopedia.com/terms/e/ enterprisevalue .asp. [14] J. Maverick,"Using EV/EBITDA and Price-to-Earnings (P/E) Ratios to Assess a Company," 05 February 2020. [Online]. Available: https://www.investopedia.com/ask/answers/072915/how-can-evebitda-be-used-conjunction-pe-ratio.asp. [15] "SPDR S&P 500 (SPY): Historical Returns," 31 March 2020. [Online]. Available: http://www.lazyportfolioetf.com/etf/spdr-sp-500-spy/. [16] S. Madge, "Predicting Stock Price Direction using Support Vector Machines," 2015. [17] T. Yiu, "Understanding Correlation And Diversification," [Online]. Available: https://towardsdatascience.com/understanding-correlation-and-diversification- 661c19a26555. 39

Abstract

Portfolio Management is the art of choosing and managing a group of financial securities such as bonds or equity instruments, derivative instruments like futures and forward contracts etc. There many key elements to portfolio management, such as asset allocation, diversification of the portfolio, and rebalancing the portfolio. In this project, my objective is to develop strategies to build optimal portfolios which aim to outperform index funds such as the S&P500 Index by researching and developing different heuristic methods for the key elements of portfolio management.

One of the strategies use various fundamental factor information such as return on an asset, asset turnover, net margin etc., to build alpha factors. The alpha factors are used as input variables in a machine learning clustering model which is used to diversify and build portfolios from the stock universe. Another strategy uses Piotroski F-scores which is calculated using financial metrics which can be grouped into three criteria's, leverage, profitability and liquidity to filter out companies with bad fundamentals. This is combined with machine learning models which predict the future movements of price and volume for the value stocks which is used to decide whether to trade long or short.

Different optimization techniques such as mean-variance optimization and minimax optimization are studied and tested on the trading strategies to allocate optimal weights to the underlying assets in the portfolios. These strategies are developed using python programming language and using the Quantopian API for data collection and back-testing in a trading environment.

Acknowledgements

I would like to express my deep sense of gratitude and thanks to my final year project supervisor, Mr. Van Ting for his constant guidance, encouragement and feedback throughout the two semesters. His guidance and recommendations played a significant role in the success of this project. His detailed feedbacks helped me to achieve better results and improve the overall quality and performance of this project.

Next, I would like to thank my mentors and managers during my various internships who introduced me to the field of data science and finance. It has been a great learning experience so far and I will continue this journey into my future career. I would also like to express my thanks to the EE department at City University of Hong Kong for providing a world class educational experience the past four years of my undergraduate studies and all the professors who have taught me during this time.

Finally, I would like to thank my parents and family for giving me the opportunity to study in Hong Kong and for the constant motivation to achieve more every day and work hard along the way. Their support has been instrumental in shaping my adulthood and making me who I am today. To say the least, I’m excited and looking forward to opening a new chapter in my life after graduation.

1. Table of Contents

List of Figures

List of Tables

1. Introduction

In the world of finance, portfolio management and optimization are a problem which is solved everyday by asset managers in the field. Portfolio management and optimization is the process of allocating capital across different financial assets such as equities, bonds and mutual funds to build investment solutions. While building a portfolio, the investor must have an investment objective and must address various factors such as how much risk they are prepared to take, the duration of their investment and many other factors which will be discussed in this paper.

Over the past decade, there has a been exponential growth and research in the field of artificial intelligence and machine learning. Advanced algorithms are being developed and tested in multiple areas and finance is one of the front runners. In this project, I will tackle portfolio optimization with help of different machine learning models and statistical models to optimally allocate weights to various equity assets and build portfolios which beat an index benchmark. The algorithms will be back tested in a real-world trading environment using Quantopian, an online quantitative research tool which provides back testing framework and also numerous APIs to get financial data. Building portfolios with the goal of beating an index benchmark is commonly known as Active Portfolio Management in the industry.

Various types of fundamental analysis and techniques for market research and company stock analysis are discussed and implemented in the strategies built. Market factors and style factors are used in the alpha discovery and research stage to build an alpha which has good predictability. Quantopian, the online quantitative analytics and backtesting platform is used to conduct research on historical datasets, alpha research and more importantly to build the trading strategies and backtesting for evaluating performance.

2. Background

2.1. Fundamentals of Investment

2.1.1. Asset

An asset is essentially anything which possesses monetary value. Some example of assets include cash, checking and savings accounts, treasury bills, stocks, bonds etc. In this project I will be focusing on stocks.

2.1.2. Stocks

A stock or equity is an asset or a form of security which indicates a proportionate ownership in a company by the holder of the stock. Stocks are generally issued by corporations to raise funds in order to operate their businesses. Stocks are bought mostly on stock exchanges. Some common and famous stock exchanges are the New York Stock Exchange (NYSE), the London Stock Exchange (LSE), NASDAQ to name a few. Stocks are usually the foundation of almost every portfolio built.

2.1.3. Portfolio

A portfolio is a set or group of publicly traded assets such as equities, bonds, mutual funds, commodities etc. Sometimes portfolios can also include non-publicly traded assets such as real estate and art. Investors construct their portfolios based on their risk tolerance their investment objectives and these portfolios are generally managed by the investors themselves or wealth managers.

2.1.4. Portfolio Construction and Optimization

Portfolio Construction is all about selecting the right assets based on the investor's objectives and risk tolerance. The most important aspect of constructing a portfolio is the asset allocation. Breaking down the portfolio optimization process, there are four important steps one must consider:

- Risk Profiling - High Risk or Low Risk based on a scale determined by the investor.
- Asset Allocation - The right combination of assets for the portfolio.
- Fine-Tuning the Portfolio - Based on the risk profile and the existing assets, add or remove the underlying assets most suitable to the investor, potentially reducing the risk and increasing the returns.

- Rebalancing - The investor must review the portfolio from time to time and rebalance the weight allocated to the underlying asset based on the market.

2.2. Fundamentals of Portfolio Theories

2.2.1. Modern Portfolio Theory

In 1952 Harry Markowitz published a paper called "Portfolio Selection" where he introduced the Modern Portfolio Theory (MPT). MPT theorizes on how a risk-averse investor can build and optimize their portfolios to achieve maximizing expected returns for a given level of market risk.

2.2.2. Portfolio Expected Returns

In a portfolio of *N* assets, the expected returns are calculated as the weighted average of the individual returns of all *N* assets. For example, if a portfolio contains 5 equally weighted assets, each with individual returns of 6%, 7%, 12%, 4% and 15% respectively, then the expected returns of the portfolio will be calculated by:

$$E(R) = (0.06 \times 0.20) + (0.07 \times 0.20) + (0.12 \times 0.2) + (0.04 \times 0.20) + (0.15 \times 0.20) = 8.8\%$$

2.2.3. Efficient Frontier

Every possible weighted combination of assets in a portfolio can plotted onto a graph showcasing the portfolio risk on the X-axis and the portfolio expected return on the Y-axis. Based on this graph an investor can choose the most desirable combination of assets. The upward sloping hyperbola which connects all the most efficient portfolio asset combinations is called the efficient frontier. For an investor following the Modern Portfolio Theory, investing in a portfolio not on the efficient frontier is not desirable.

2.2.4. Sharpe Ratio

Nobel laureate William Sharpe introduced a new method to understand a portfolio's performance. The Sharpe Ratio is used by investors to understand the return on investment (ROI) compared to the amount of risk undertaken. The Sharpe ratio can be calculated by using the formula:

$$SharpeRatio = \frac{R_p - R_f}{\sigma_p}$$

Where:

$$R_p = ExpectedReturnofPortfolio$$
$$R_f = RiskFreeRate$$
$$\sigma_p = Standard\ Deviation\ of\ Portfolio's\ Excess\ Returns$$

2.3. Active Portfolio Management

In the industry there two main types of investment management strategies practiced by portfolio managers, passive portfolio management and active portfolio management. Passive investing generally involves replicating the security holdings of an index fund to achieve similar results. Active investing on the other hand involves managing portfolios in order to beat an index or a benchmark to achieve superior returns. As the name suggests, in active investing a portfolio manager trades more frequently and conducts more in-depth fundamental research. Active portfolio management does involve high risk compared passive. In my project my goal is to implement active portfolio management and build strategies aimed to beat the Standard & Poor's 500 index [1].

2.4. Fundamental Factor Research [2]

Financial factors are essentially the foundations of active investing. They refer to various ratios and metrics which are used to measure a company's financial characteristics which are derived from financial statements, quarterly reports, balance sheets, etc. Some examples of such factors are market capitalization, net income, net margin and cash flow. I will make use of similar fundamental factors in one of strategies that help to determine characteristics that affect the portfolio's assets risk and return. One of the important steps is to select the significant factors which can be done by conducting in-depth factor research [2] [3] [4].

2.5. Factor Based Strategies

With respect to investing strategies, there are two main types of factors that have driven returns, macroeconomic factors and style factors. Some common style factors used in portfolios include:

a. Momentum: This factor identifies stocks which have performed well in the past and tends to exhibit high returns in the future.
b. Value: This factor identifies stocks which have low prices when compared to their fundamental factors. This is used in determining different metrics like the Piotroski score explained ahead.
c. Volatility: This factor is essentially the standard deviation or risk of a particular asset. Various research studies have shown that stocks with lower volatilities earn higher risk adjusted returns.
d. Size: This is one of the most important factors which can used to diversify a portfolio. The size of an asset is captured by the market capitalization. Historically, small-cap stocks exhibit higher returns compared to large-cap stocks.

These factors are used to help improve portfolio outcomes, reduce the portfolio volatilities and diversify the portfolio assets [4]. I make use of style factors in my algorithm as input variables for my machine learning model.

2.6. Piotroski Score

The Piotroski F-Score is a score used to determine the strength of company based on their financial statements. This score is used to filter out stocks of weak companies and determine the top value companies. The score is awarded from zero to nine based on a few criterions where zero is the worst and nine is the best. The criterions used to calculate the score are as follows [5]:

Profitability (4 Points):

- Positive net income.
- Positive return on the underlying asset.
- Positive cash flow.
- Cash flow being greater than net income.

Leverage and Liquidity (3 Points):

- Decreased leverage compared to previous year.
- More liquidity compared to previous year.
- No new shares issued in the previous year.

Operational Efficiency (Two Points):

- Increased gross margin compared to previous year.
- Increased asset turnover compared to previous year.

Based on the above criterions, a stock universe can be ranked and the top ranking stocks can be selected for further diversification or constructing a portfolio [5]. This method will be further explained in the methodology and implementation section of the report.

2.7. Machine Learning Models

2.7.1 K-Means Clustering

Clustering analysis is an unsupervised machine learning method which is used for finding homogenous groups from the data in a way that the objects in each group share similar characteristics to each other when compared to other group objects. There are many mathematical clustering algorithms designed which have different definitions of how clusters are calculated such similarity or distance functions like Euclidean distance or correlation distance and how many number of clusters are created.

Clustering algorithms can be divided into two main types of algorithms: partitional clustering and hierarchical clustering. For my project I will be implementing partitional clustering algorithms which creates partitions in the data by grouping disjoint clusters since there is no hierarchical nature to stock data. K-Means Clustering algorithm is a commonly used partitional algorithm where it groups ***n*** observations into ***k*** number of clusters where each observation belongs to the cluster centroid with the nearest mean.

The objective function of the algorithm below is to be minimized which is the cluster sum of squares:

$$arg_s \min \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Where x are the number of observations in the data, $S = S_1, S_2, \ldots, S_k$ are the set of observations and μ_i is the mean of each points in S_i

K-Means is an iterative algorithm which selects the centroids of each group randomly in the first iteration which are used as the initial points for each cluster. From here it optimises the positions of the centroid with repetitive calculations. The approach for calculating the positions in every iteration is called Expectation-Maximization and can be mathematically defined by the equation above. The figure 1 below shows clustering example of stock returns data where we have three clusters and each cluster with a centroid.

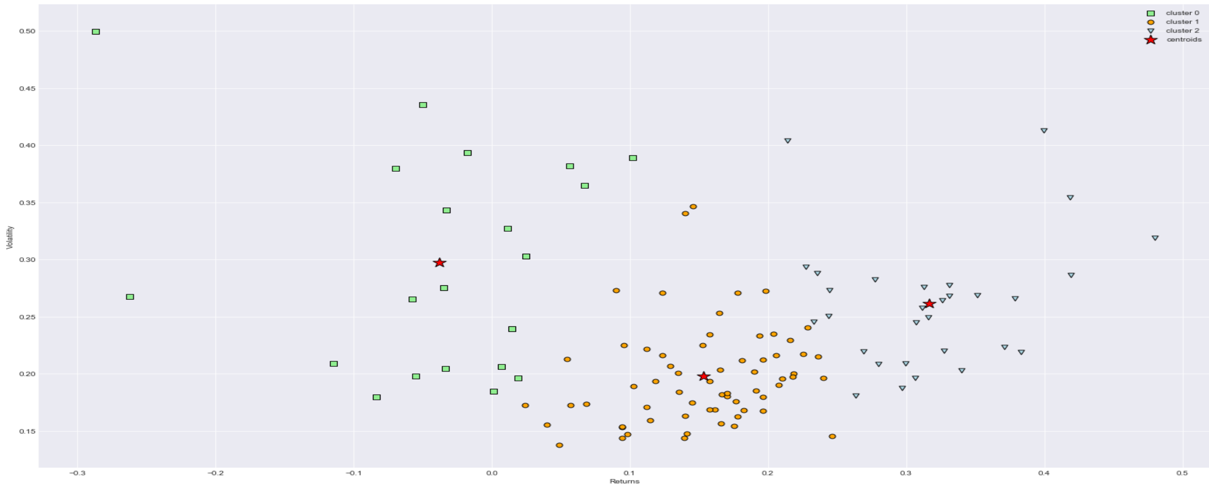

Figure 1 - KMeans Clustering example with centroids

The algorithm is complete when the number of iterations are completed or when the centroids have no change in their values, meaning they are stabilised. This entire process can be shown graphically by the below diagram figure 2.

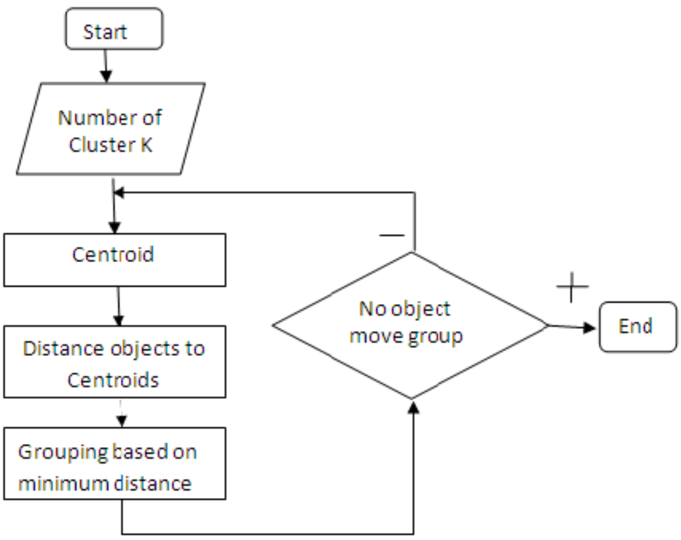

Figure 2 - KMeans Clustering Flowchart https://www.researchgate.net/figure/K-Means-Clustering-Flow-Chart_fig1_269800441

2.7.2. Support Vector Machines

Support vector machines is a machine learning model which is used for classification and regression tasks but more widely used for classification. The objective function of a support vector machine is to calculate the hyperplane in an N-dimensional space where N is the number of input features. The hyperplane is calculated in an iterative manner in order to minimize the error [6].

The three main concepts in an SVM algorithm are:

- Margin - This is the gap between the closest datapoints of two separate classes. The objective of the algorithm is to maximise the margin by finding the largest distance.
- Support Vectors - Support Vectors are datapoints that are closer to the hyperplane and influence the position of the hyperplane. They are used to maximise the margin of the classifier.

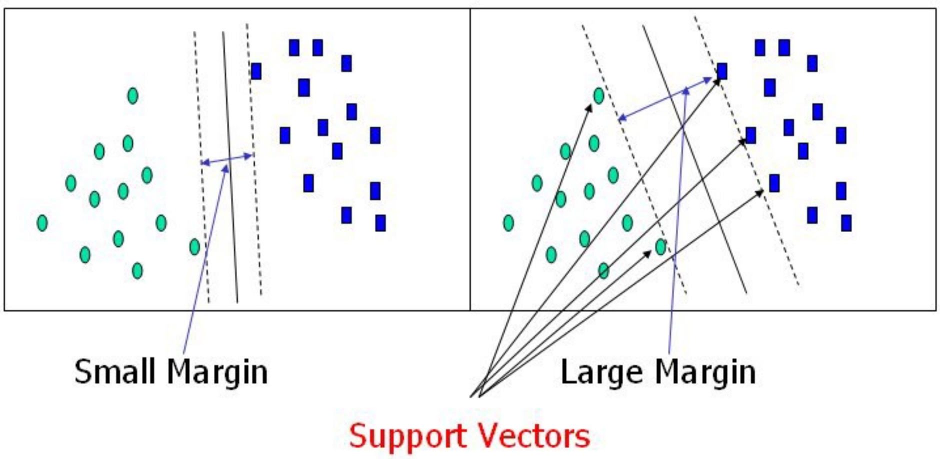

Figure 3 - SVM Margin and Vectors

- Hyperplanes - They are the decision boundaries which help to distinguish between datapoints and classify them accurately. Dimensions of hyperplanes are dependent on the number of input features used to train the SVM.

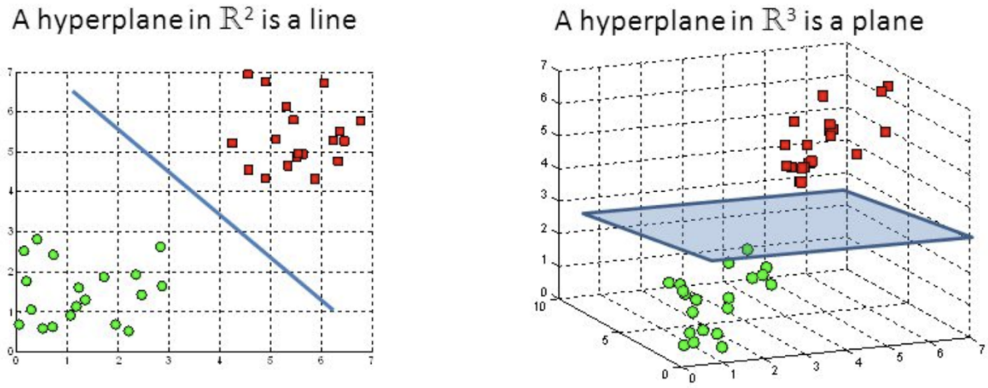

Figure 4 - SVM Hyperplanes in 2D and 3D

2.7.3. Gradient Boosting Regressors

Gradient Boosting algorithm is another type of machine learning technique which is used for both classification and regression problems. Gradient Boosting Regressors (GBR) is a form of ensemble tree model, generally decision trees. The reasons for difference between the actual and predicted values when using a machine learning technique is mainly due to noise, variance and bias. With the help of ensemble models like GBR, these factors can be reduced [7].

Boosting is an ensemble technique where the predictors are made sequentially where subsequent predictors learn from the mistakes of the previous predictors. This helps to reduce the bias and variance. The intuition behind the GBR algorithm is to repetitively leverage the patterns in residuals and strengthen the model with weak predictions and make it better [7].

The objective of the GBR algorithm is to minimize the defined loss function. Loss functions are defined based on the problem being solved and for a regression task the loss function can be the mean squared error [7].

2.8. Portfolio Optimization Techniques

2.8.1. Mean Variance Optimization

Mean-Variance optimization technique was first introduced in Harry Markowitz's Nobel Prize winning work on 'Modern Portfolio Theory'. An investor who follows the MPT for their portfolio, models the rate of return on the underlying assets as a random variable [8]. The objective of the optimization is to choose the weights for the assets in the most optimal method. Mean Variance optimization uses standard deviation of a portfolio as the measure of risk. The inputs for the optimization model are as follows:

- The expected returns for each asset in the portfolio.
- The standard deviation (risk) of each asset in the portfolio.
- The correlation matrix between each asset in the portfolio.

The output of the optimizer is called the Efficient Frontier. The efficient frontier can be visualized on a graph which shows a set of optimal portfolios from which an investor can choose the portfolio which offers the highest expected returns for a given level of risk or the lowest risk for a given level of returns [8].

The optimization algorithm for a portfolio of two securities ***j*** and ***k*** with data points $\{y_{jt}, y_{kt}\}$ can be mathematically denoted as:

$$min_w \sum_{j=1}^{N} \sum_{k=1}^{N} w_j w_k s_{jk}$$

Parameter Constraints:

$$\sum_{j=1}^{N} w_j \bar{y}_j \geq G$$

Where G is the minimum level of expected returns of the portfolio.

$$\sum_{j=1}^{N} w_j < W$$

Where W is the maximum weight allocation available which is 1 (100%).

$$S_{jk} = \frac{1}{T-N} \sum_{t=1}^{T} (y_{jt} - \bar{y}_j)(y_{kt} - \bar{y}_k)$$

Where S_{jk} is the covariance matrix between the two securities.

2.8.2. Mini-Max Optimization

Mini-Max optimization is a simple linear programming algorithm which chooses the optimal portfolio based on the minimum return rather than minimum variance like the mean-variance optimization technique. Being a linear programming algorithm, Mini-Max makes it possible to constrain some variables as Boolean 1-0 or integer values. Other than the computational convenience, it also has logical advantages when the returns of the assets are non-normally distributed [8].

Suppose we have a portfolio with N assets over a time T, let:

$$y_{jt} = Return\ on\ asset\ j\ at\ a\ particular\ time\ t$$

$$\bar{y}_j = Mean\ of\ returns\ of\ asset\ j$$

$$w_j = Weight\ allocated\ to\ asset\ j$$

$$y_{pt} = \sum_{j=1}^{N} w_j y_{jt} = Return\ on\ portfolio\ at\ time\ t$$

$$E_p = \sum_{j=1}^{N} w_j \bar{y}_j = Average\ Return\ on\ portfolio$$

$M_p = Minimum\ Return\ on\ portfolio$

The optimization maximizes the minimum return M_p , that is it minimizes the maximum loss where loss is the negative gain, in other words maximize the minimum gain. This can be derived as follows:

$max\ M_p\ \ Subject\ to\ the\ following\ constraints$:

$$\sum_{j=1}^{N} w_j y_{jt} - M_p \geq 0, \qquad t = 1, 2, \dots, T$$

$$\sum_{j=1}^{N} w_j \bar{y}_j \geq G \quad where\ G\ is\ the\ minimum\ expected\ returns$$

$$\sum_{j=1}^{N} w_j \leq W \quad where\ W\ is\ the\ maximum\ weight\ allocation$$

2.9. Quantopian Environment

Quantopian is an open source quantitative research and analytics tool which provides numerous APIs and a back testing environment for trading algorithms. For my project I will be using Quantopian mainly for alpha research and back testing since it makes it easier to access good quality financial data through their API and also they provide more computational power than compared to my local machine since it is hosted online.

3. Objective

For my final year project, my objective is to learn the fundamentals of quantitative finance and machine learning and apply my knowledge into building trading strategies which can beat the benchmark index which is the Standard & Poor's 500 index. Over the course of the project I built and tested various strategies trying to beat the benchmark. I have finally chosen two strategies to present which not only beat the benchmark but also achieved a relatively high Sharpe ratio.

4. Literature Review and Related Work

4.1. Portfolio Diversification using Cluster Analysis

A lot of the research work previously achieved in this field has received positive sentiments from the industry experts. Correlation is a fundamental statistical method which gives us the likelihood that two variables move together. A positive correlated pair essentially means when one goes up the other goes up too or vice versa. A negative correlated pair is the opposite, when one moves up the other moves down. Applying this to the stock market, investors try to build portfolios with negative correlated assets so that they can diversify their portfolios and are less subjected to market risks. *Correlation based clustering of the Stockholm Stock Exchange by Frederick Rosen (2006)* is good example of a research paper where the author classifies stocks based only on the correlation between them. In this method of clustering if one clusters stock price decreases, another clusters stock price will not decrease thereby hedging to reduce loss. Correlations between stocks is not the most efficient way to diversify a portfolio since they often reverse or change during periods of economic stress. Therefore during the times of economic stress the assets within a cluster may not be highly correlated with each other or clusters may not be negatively correlated with each other. This is one drawback from clustering based on correlations and defeats the purpose of diversification from an investor standpoint by exposing the portfolios to higher risk [9].

4.2. Piotroski F-Score to identify value stocks

As I mentioned in the background study, Piotroski F-Score is used to identify high value stocks and reduce an investors stock universe. There have been various researchers who have developed and tested numerous methodologies to improve the returns of a portfolio in the value universe. Joseph Piotroski (2000) tested his investment strategy between the time periods 1976 and 1996 using American stocks. Piotroski calculated the book-to-market at the end of the financial year for each firm and assign them to a quintile. He then calculated the F-score for each firm in the highest quintile assigning them a score between 0-9 where 9 is the highest value firm. He created two portfolios, one containing low F-score stocks and other containing high F-score stocks. He documented that the high F-score portfolio outperformed the low F-score portfolio [10]. One of the most useful benefits of using this method devised by Piotroski is that this strategy helps in identifying small and mid-capitalization companies which have high value. These companies generally have low fundamental analysis coverage and neglected by investors [11].

4.3. Minimax and Mean Variance Optimization

There are many portfolio optimization techniques used in the industry. Markowitz (1952) introduced the mean-variance optimization technique in his Nobel prize winning essay on Modern Portfolio Theory. *A Minimax Portfolio Selection Rule with Linear Programming* by *Martin Young (1996)* introduces a new optimization technique which uses historical returns data similar to mean-variance and is a solution to a linear programming problem. This optimization technique uses the minimum returns of an asset as the measure portfolio volatility as compared to the mean-variance optimization which uses the variance [8].

5. Strategy Design and Implementation

5.1. K-Means Clustering with MiniMax Optimization (Strategy 1)

This is the first trading strategy that I developed in which I implement K-Means clustering for portfolio diversification and use the Mini-Max optimization algorithm for optimal weight allocation.

5.1.1. Alpha Research

In this strategy I use four different factors equi-weighted into one alpha factor which is used as the input feature for the K-Mean clustering algorithm. Choosing the right factors requires a lot of research to understand how the factors behave and affect a portfolio. Quantopian has a research environment where we can use a tool called ‘AlphaLens’ for analysing a factor’s effectiveness at predicting future returns. I can also create visualizations which help to further understand a factor.

For my algorithm I researched and decided to the following four factors as my input feature:

a. Return on an asset = Net income / Total assets
b. Asset Turnover = Net Revenue / Total assets
c. Net Margin
d. Growth Score

Once the factors are decided, I combine them together with equal weights to form one alpha in order to analyse and check if our factor data affects prices in the future. The time periods checked are 1 day, 5 days and 10 days, denoted as 1D, 5D and 10D respectively in the visualizations. The stock universe is split in quantiles where quantile 1 contains the 20% of the stocks with the lowest alpha factor values and quantile 5 with the highest 20%. I then research the returns generated by each quantile over a time period. The goal is to make sure that quantile 5 performs the best while quantile 1 is the worst.

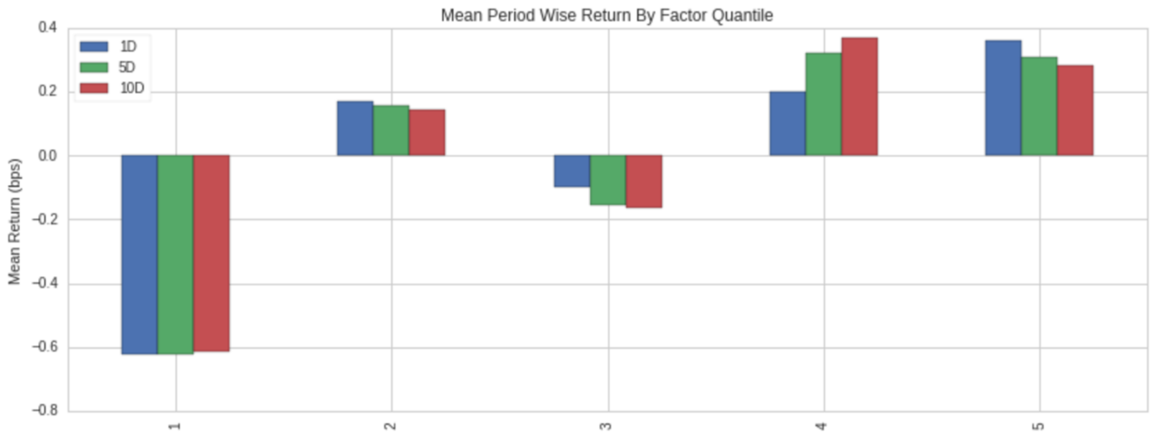

Figure 5 - Mean Period Wise Returns for Factor Quantiles

Figure 5 shows the mean returns for each quantile for the time period mentioned before. We can see that the first quantile performance is negative while the fifth quantile has a positive mean return. This is one of the many analysis we can conduct to see whether our alpha factors are good for our portfolios.

Figures 6, 7 and 8 plot the log cumulative returns of the each of the quantiles over a five year back testing period from 2015 to 2019. The visualization shows that the fifth quantile's total cumulative returns is much higher and outperforms the first quantile.

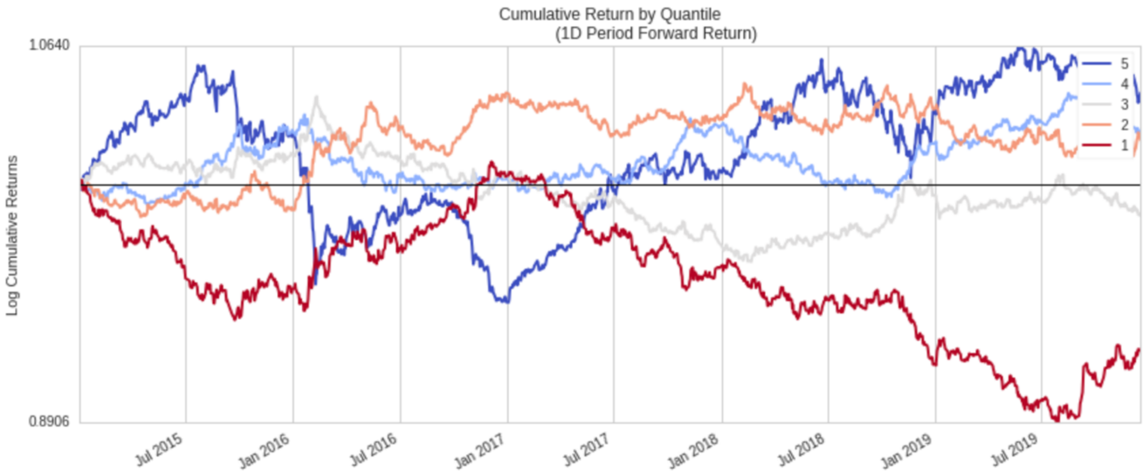

Figure 6 - Cumulative Factor Returns for 1D Forward Returns

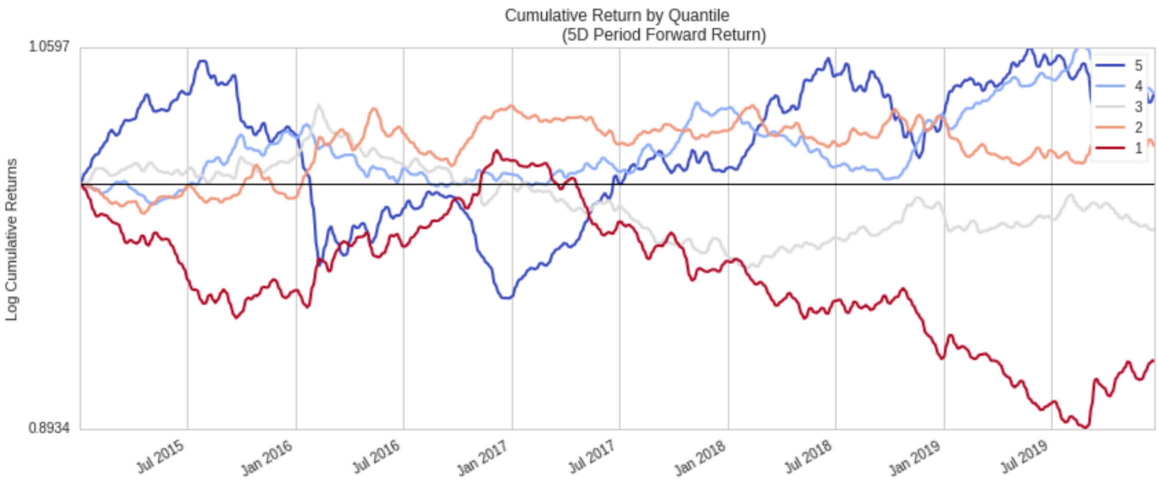

Figure 7 - Cumulative Factor Returns for 5D Forward Returns

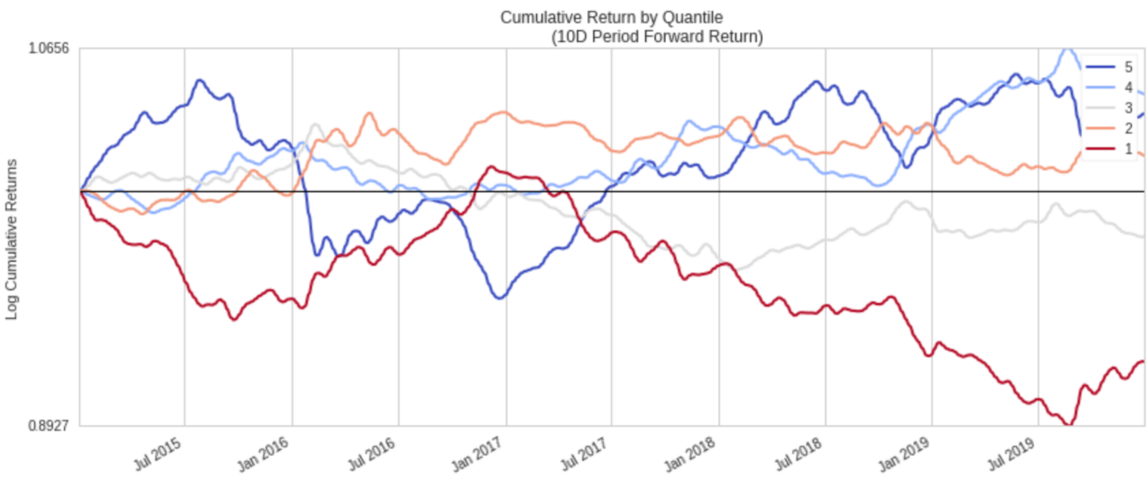

Figure 8 - Cumulative Factor Returns for 10D Forward Returns

Another important metric to determine our alpha factor performance is the information coefficient (IC). The IC is used to describe the correlation between the predicted and actual returns and sometimes also evaluate the skill of a portfolio manager or quantitative investment analyst. The score is between -1 to 1 and quantifies the predictiveness of the alpha factor. Figure 8 shows the IC decay for the alpha factor over the period of one trading year (252 days). The alpha factor is considered useless when the IC drops below 0. However, in my alpha factor, we can see the IC score gradually increase over time.

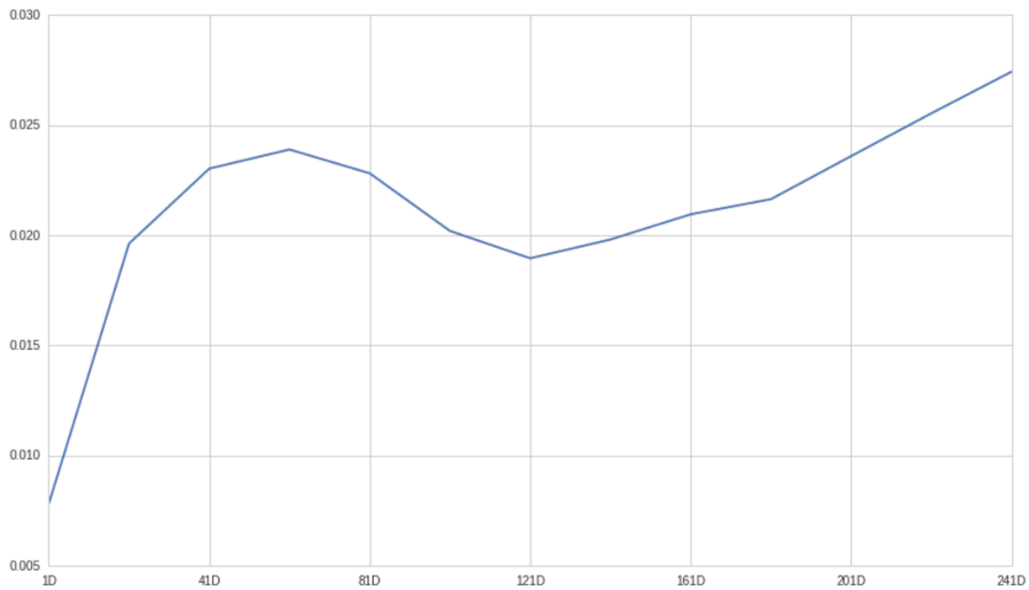

Figure 9 - Information Coefficient Decay for Alpha Factor

5.1.2. Data Collection

The factor data, stock universe data, stock prices data are all collected from a Quantopian API. This makes it data collection faster and much simpler in terms of data pre-processing and cleaning. After the alpha research phase, I started the data collection for the factors and building the foundation for my trading strategy. The stock universe used for my strategy is the Quantopian 1500 stock universe which provides a list of the top 1500 tradeable equities where tradeable can be defined as:

Stock must be the primary share class for its company.

- The company must have a known market capitalization.
- The stock must not be a depository receipt.
- The stock must not be traded over the counter.

- The stock must have a previous closing day price.
- The stock must have a volume greater than zero on the previous trading day.

With the stock universe available to choose from, I add another parameter to filter only the companies issuing stocks with a market capitalization of greater than $1.5 billion USD which is the median market cap for the S&P 600 index so that I can include small, mid and large market cap companies in my universe.

5.1.3. Clustering Algorithm

Once the factor data is collected, the clustering function creates an equi-weighted alpha factor similar to what was done in the research stage. This alpha factor is used as the input feature for the KMeans clustering function. The KMeans function has the hyperparameters set to n_clusters = 15, n_jobs = 1 and random_state = 10. The algorithm outputs a cluster for each asset in the stock universe and all single stock clusters are removed if there are any. Once the clustering algorithm is complete, high performing stocks are filtered from each cluster by calculating the Sharpe ratio of all the stocks by taking the rolling price data of three-month period. The highest Sharpe ratio stock from each cluster are selected for the portfolio.

5.1.4. Weight Optimization

Once the stock universe is selected, I calculate the optimal weights for each asset in the portfolio by using the Mini-Max optimization technique previously discussed. The optimizer has maximum weight of 5% per asset as one of the constraints set since I don’t want to have a high exposure for any single asset. The optimization algorithm uses the historical daily returns for a two-month period as one of the inputs and outputs the target weights for each asset.

5.1.5. Portfolio Rebalancing and Trading

The portfolio is rebalanced every month by calling the clustering and weight optimization functions. Depending on the target weights, an asset is traded long or short. The initial investment for trading is $1,000,000 USD. The rebalancing takes place three minutes after the market opens on the first day of each month.

5.2. Piotroski F-Score value stocks (Strategy 2)

This is the second strategy that I backtested which beat the benchmark index. In this strategy I filter high value companies using the Piotroski f-score method and then predict whether to go long or short using a machine learning algorithm and historical data. The weights are optimized using mean-variance optimization.

5.2.1. Data Collection

Similar to the previous algorithm, the stock universe used is the Quantopian 1500 stocks list. To improve the performance of the algorithm, I conduct a few stock valuations analyses on the stock universe and exclude stocks which don't meet the below selection criteria's:

a. Companies with market capitalization greater than $1Billion USD
b. Companies with positive earnings before interest and tax (EBIT). This metric indicates a company's profitability and is generally used to analyze the performance of a company's operations [12].
c. Companies with positive enterprise value (EV). This metric is a measure of a company's total value where EV = Market Cap + Total Debt - Cash (liquid assets). Enterprise value is used as a foundation for many financial ratios used to measure performance [13].
d. Companies with positive EV/EBITDA ratio. This metric compares a company's enterprise value to the earnings before tax, interest, depreciation and amortization. This ratio helps in identifying undervalued companies [14].
e. Companies with low annualized volatility stocks. Selecting stocks with low volatility helps in reducing the beta of the overall strategy performance.

Once the universe is defined, the Piotroski f-score is calculated for each stock in the universe. The pipeline returns the stocks and their respective scores.

5.2.2. Machine Learning Model

The top fifty stocks with the highest Piotroski score are selected for the portfolio rebalancing every month. Two machine learning models are trained using the historical data of each stock. The first model is used to train the historical pricing data of a stock so that the present data can be used to predict whether a stock price will be increasing or decreasing. This information can be used to decide whether to go long or short the stock. Similarly, the second model uses the historical volume data of stocks to predict the change in volume. The historic range used as training data is set to 180 days (6 months). For the prediction data, the lookback period is 2 days since the latest pricing and volume data is important to predict the future movements. The machine learning model used is the Gradient Boosting Regression algorithm discussed before.

5.2.3. Weight Optimization

Once the machine learning model predicts the price movement of the stock and whether to go long or short, the optimal weights are calculated using mean-variance optimization previously discussed in detail. The optimization algorithm uses the historic two-month pricing data to calculate the optimal weights for the portfolio.

5.2.4. Portfolio Rebalancing and Trading

The rebalancing period is set to the beginning of each month, three minutes after the market opens similar to the previous strategy. The initial investment used for the algorithm is $1,000,000 USD. If both the machine learning models predictions are greater than zero then the stock is traded long with the optimal weight, if both the predictions are lesser than zero then the stock is traded short and finally if one prediction is greater than zero and the other lesser, then the stock is not traded or rebalanced for that month.

6. Results and Performance

6.1. K-Means Clustering with MiniMax Optimization (Strategy 1)

The strategy was back tested for a period of 5 years from 2015-01-01 up until 2019-12-31. From Table 1 below, the overall performance of the strategy during the backtest period can be deduced.

Table 1 - Strategy Performance metrics (Strategy 1)

Performance Metrics	Backtest
Annual Returns	16.828%
Cumulative Returns	117.233%
Annual Volatility	12.574%
Sharpe Ratio	1.30
Sortino Ratio	2.00
Maximum Drawdown	-13.645%
Daily Turnover	7.962%
Alpha	0.12
Beta	0.39

The cumulative returns of the strategy are 117.233% compared to the benchmark S&P500 Index returns of 72.82% during the time period. The strategy also has an alpha score of 0.12 or 12% which is the excess return compared to the benchmark and risk undertaken. The overall Sharpe Ratio is 1.30 which is considered to be good by investors.

It's also good to understand how the strategy is performing with respect to factors such as size, momentum, value and also for each sector. This is useful so that the strategies can be optimized further to get better results.

Table 2 - Performance related to common risk (Strategy 1)

Summary Statistics	
Annualized Specific Return	8.36%
Annualized Common Return	7.59%
Annualized Total Return	16.83%
Specific Sharpe Ratio	0.77

Table 2 - Exposure Summary (Strategy 1)

Exposures Summary	Average Risk Factor Exposure	Annualized Return	Cumulative Return
basic_materials	0.02	0.50%	2.50%
consumer_cyclical	0.10	1.80%	9.29%
financial_services	0.05	1.08%	5.50%
real_estate	0.05	0.03%	0.17%
consumer_defensive	0.05	-0.52%	-2.56%
health_care	0.09	0.18%	0.90%
utilities	0.03	-0.00%	-0.00%
communication_services	0.01	-0.01%	-0.03%
energy	0.02	0.47%	2.39%
industrials	0.07	0.27%	1.36%
technology	0.07	0.93%	4.74%
momentum	0.66	1.05%	5.36%
size	0.39	0.32%	1.62%
value	-0.28	0.08%	0.40%
short_term_reversal	-0.34	0.20%	1.02%
volatility	-0.24	1.11%	5.66%

Based on the cumulative returns performance for the 11 sectors and the 5 style factors seen on table 2, I plot the returns of the factors and sectors on a graph as seen in the figures 10 and 11 below. Momentum and volatility are the best performing style factors while sectors such as consumer cyclical which includes industries like automobile, housing, entertainment and retail, financial services sector and technology sector stocks generate superior returns.

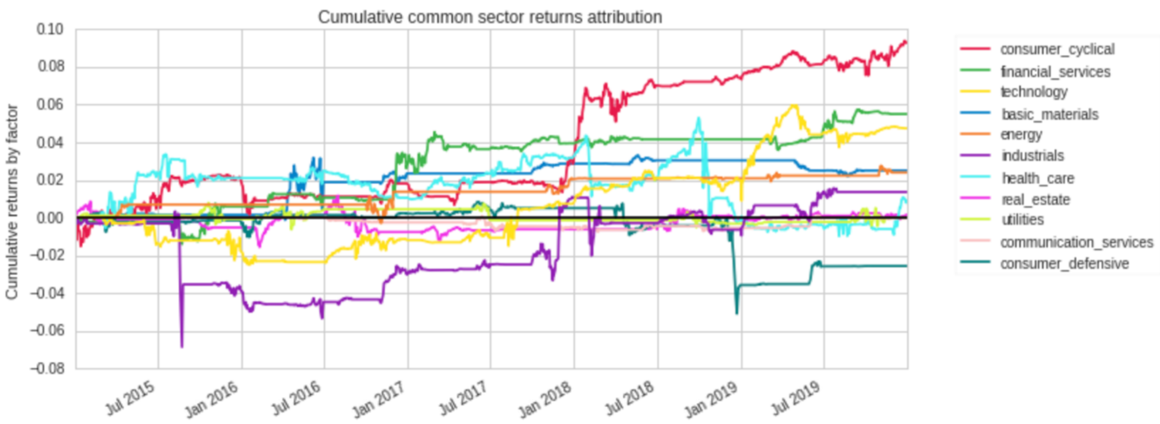

Figure 10 - Cumulative returns for sectors (Strategy 1)

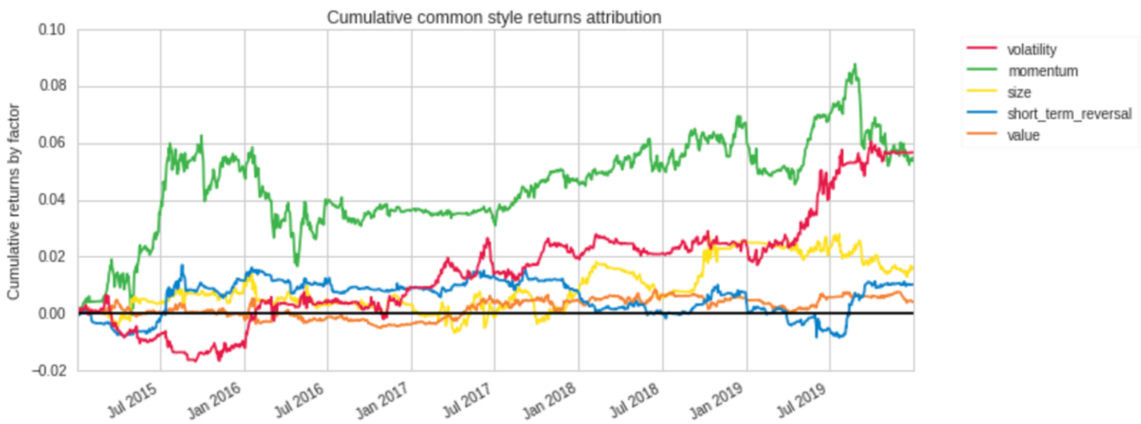

Figure 11 - Cumulative Returns for Style Factors (Strategy 1)

Figure 12 below plots the rolling Sharpe ratio over the period of the backtest. From the graph one can understand the performance of the algorithm at various points in time. Analysing the plot, we see that the strategy performs very well during 2016 with the Sharpe ratio peaking, however during early 2019 all the way to late 2019 we see a drop in the Sharpe ratio. This may be due to exposure to various risk factors which can be further analysed to improve the algorithm.

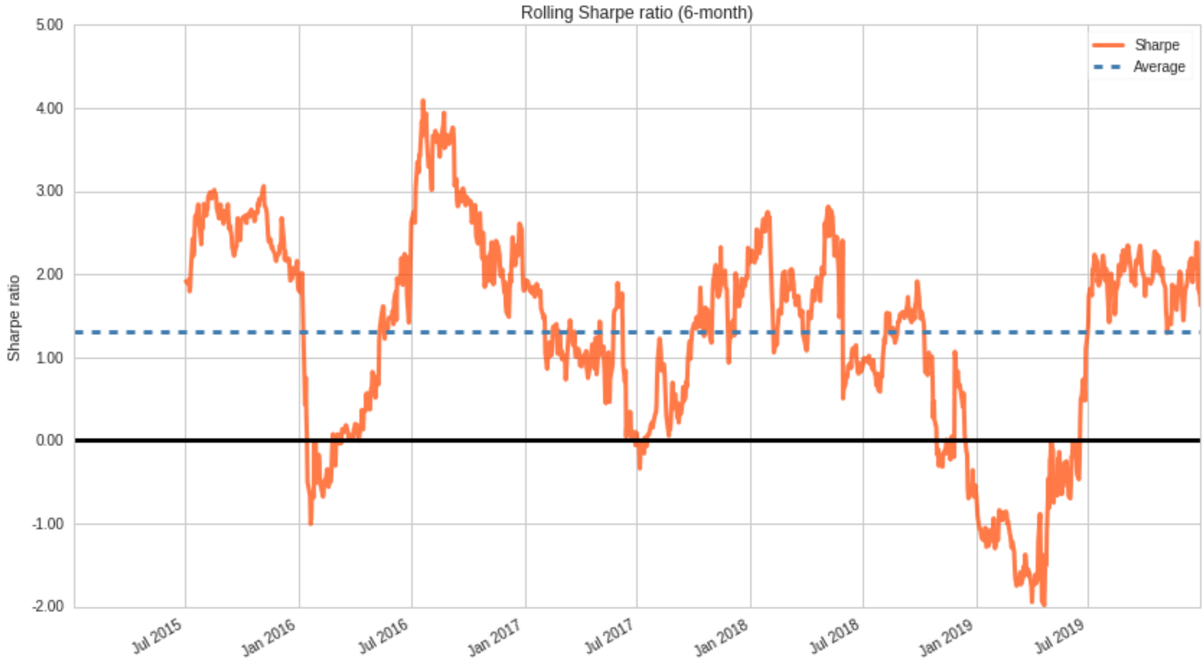

Figure 12 - Rolling Sharpe Ratio (Strategy 1)

Based on the understanding of the strategies Sharpe ratio over the time period, the annual returns graph per year in figure 14 validates the assumption made above. The highest annual returns are recorded in the year 2016 and lowest in the year 2018. Figure 13 plots the cumulative total returns along with the common returns and the specific returns which is the difference between the total returns and the common returns.

Figure 13 - Total, Specific and Common Returns (Strategy 1)

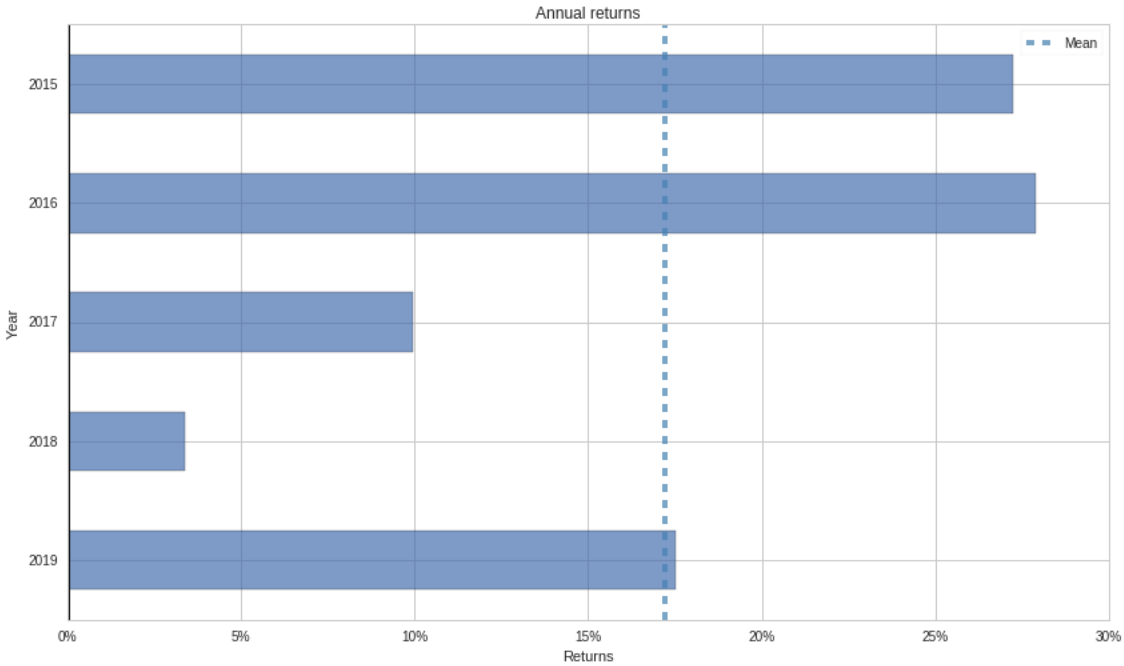

Figure 14 - Annual Returns per year (Strategy 1)

Figure 13 - Strategy Performance against Benchmark (Strategy 1)

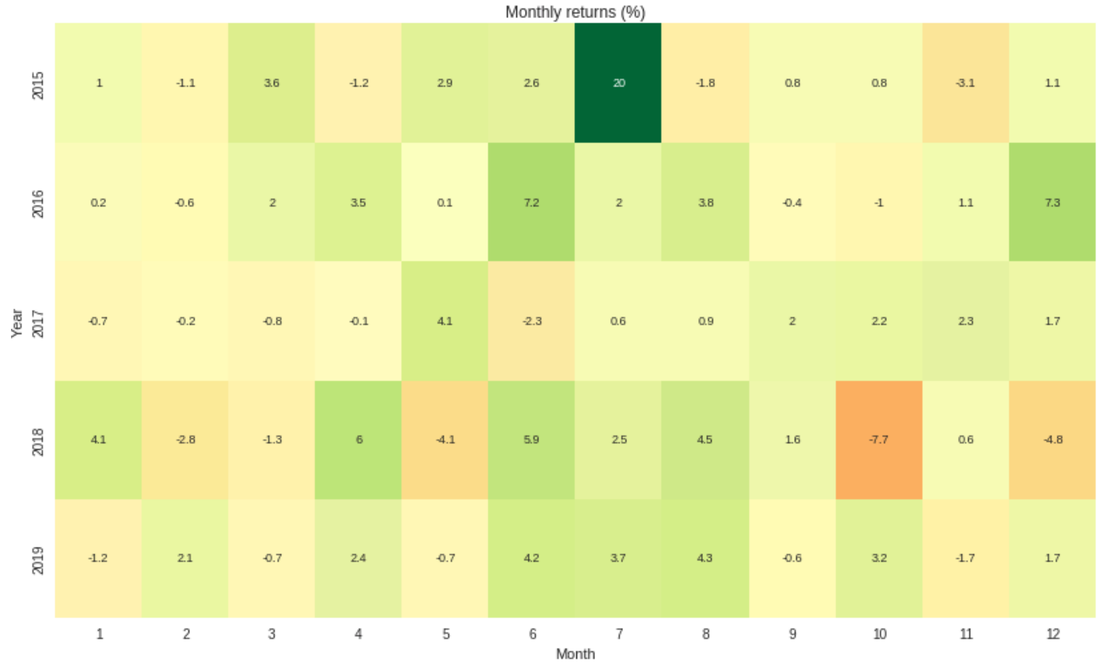

Figure 14 - Monthly performance Heat Map (Strategy 1)

Figure 13 and 14 are visualization illustrating the strategy's returns which is helpful in gauging how the algorithm performs in different time periods throughout the backtest. It also helps in identifying seasonal patterns if any and optimize the strategy. In figure 14, the top performing month of the strategy can be identified as July 2015 with 20% returns.

The drawdown of the strategy is calculated to be -13.6% over the 5 year backtest period. This means the peak to trough decline of the strategy was 13.6%. To put this into perspective, during the 2008 financial crisis, the market drawdown was over 50%. Drawdown is a good metric to measure the financial risk of a strategy. According to Quantopian a good benchmark for maximum drawdown is less than 20%. Figure 14 shows the top 10 drawdown periods for the strategy how it performed during the drawdowns and Figure 16 is an underwater plot to get a better visualization into the drawdowns. Table 3 gives more information about the drawdowns like the duration in number of days and the net drawdown percentage for each period.

Table 3 - Top 5 worst drawdowns (Strategy 1)

Worst drawdown periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	13.64	2018-09-28	2019-01-10	2019-08-09	226
1	10.48	2018-05-07	2018-05-31	2018-08-27	81
2	6.63	2015-07-31	2016-01-20	2016-04-11	182
3	6.08	2018-01-26	2018-02-09	2018-04-17	58
4	4.85	2015-03-20	2015-05-12	2015-06-11	60

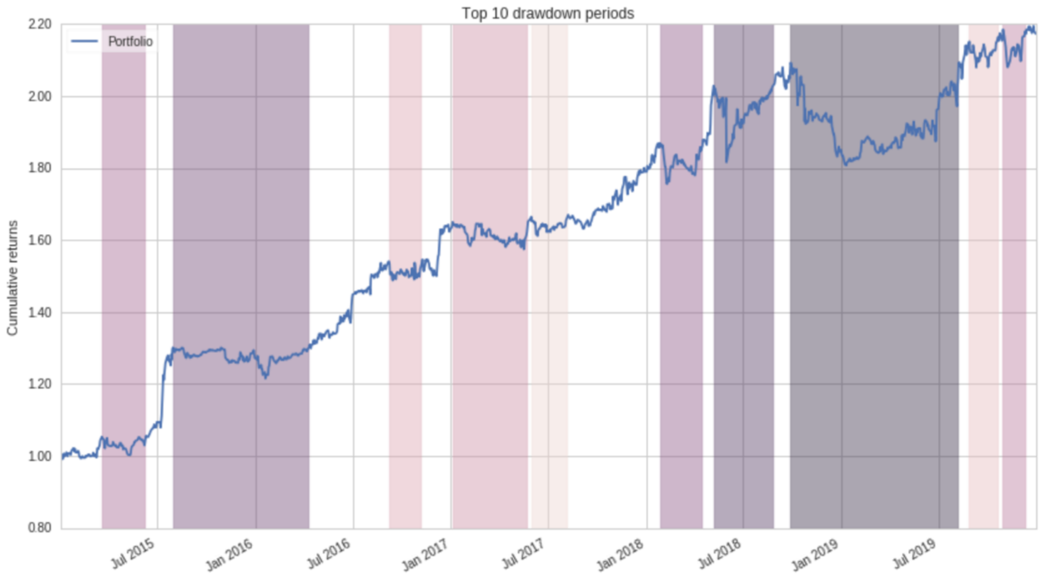

Figure 15 - Top 10 drawdowns (Strategy 1)

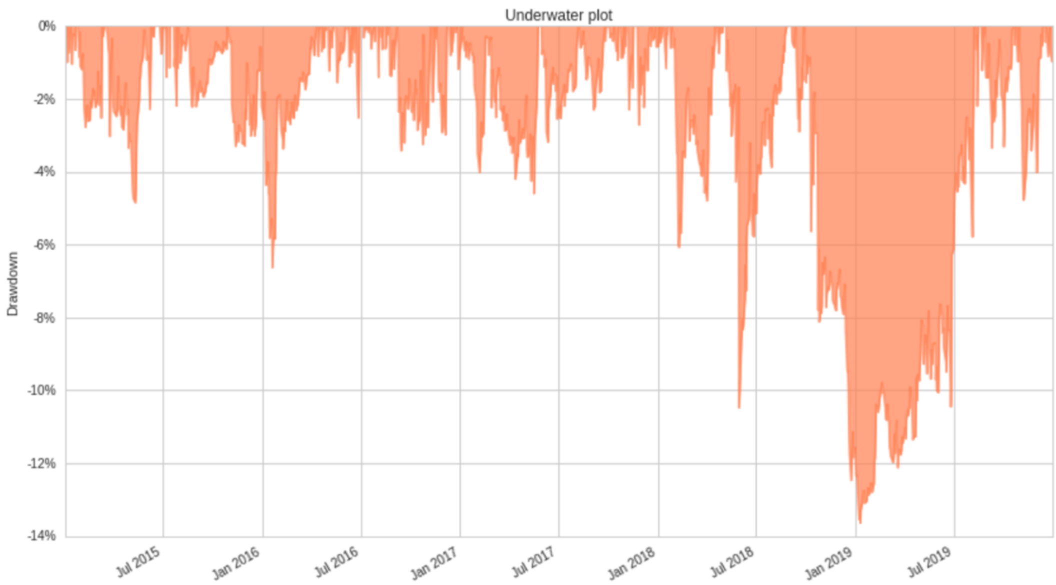

Figure 16 - Underwater Drawdown (Strategy 1)

6.2. Piotroski F-Score value stocks with Machine Learning (Strategy 2)

This strategy is backtested similarly to strategy one on the Quantopian environment for a four-year period from 2015-01-01 to 2018-12-31. The backtest performance for the strategy can be found the table below with a comparison with strategy 1.

Table 4 - Strategy 2 Performance Metrics

Performance Metrics	Backtest
Annual Returns	31.853%
Cumulative Returns	201.254%
Annual Volatility	20.291%
Sharpe Ratio	1.46
Sortino Ratio	2.40
Maximum Drawdown	-16.554%
Daily Turnover	2.245%
Alpha	0.28
Beta	0.24

Looking at the comparison column, the strategy 2 annual returns is almost twice as strategy 1 with little higher annual volatility. This explains the higher Sharpe ratio as well. Strategy 2 experiences a similar maximum drawdown, a higher alpha score and a lower beta score which are all signs of a better performance.

Table 4 and 5 is the performance relative to common risk factors and also exposures to style factors and sector information. Figure 16 plots the cumulative returns, specific returns and the common returns.

Table 5 - Performance related to common risk (Strategy 2)

Summary Statistics	
Annualized Specific Return	28.78%
Annualized Common Return	2.14%
Annualized Total Return	31.81%
Specific Sharpe Ratio	1.38

Table 6 - Exposures to Sector and Style Factors and Common Returns (Strategy 2)

Exposures Summary	Average Risk Factor Exposure	Annualized Return	Cumulative Return
basic_materials	-0.06	0.07%	0.28%
consumer_cyclical	-0.06	-1.03%	-4.04%
financial_services	0.05	0.78%	3.15%
real_estate	0.00	0.00%	0.00%
consumer_defensive	-0.09	0.45%	1.80%
health_care	0.15	0.46%	1.84%
utilities	-0.02	-0.13%	-0.52%
communication_services	-0.01	-0.30%	-1.18%
energy	0.03	-0.26%	-1.01%
industrials	-0.02	-0.20%	-0.81%
technology	0.20	3.36%	14.09%
momentum	0.32	-0.48%	-1.91%
size	-0.18	0.01%	0.04%
value	-0.24	-0.04%	-0.15%
short_term_reversal	-0.25	-1.17%	-4.60%
volatility	0.02	0.68%	2.72%

Table 6 shows that the technology sector and the financial services sector generate high positive returns and the volatility style factor has the highest returns out of the five factors while consumer cyclical sector stocks and short-term reversal factor have negative cumulative returns.

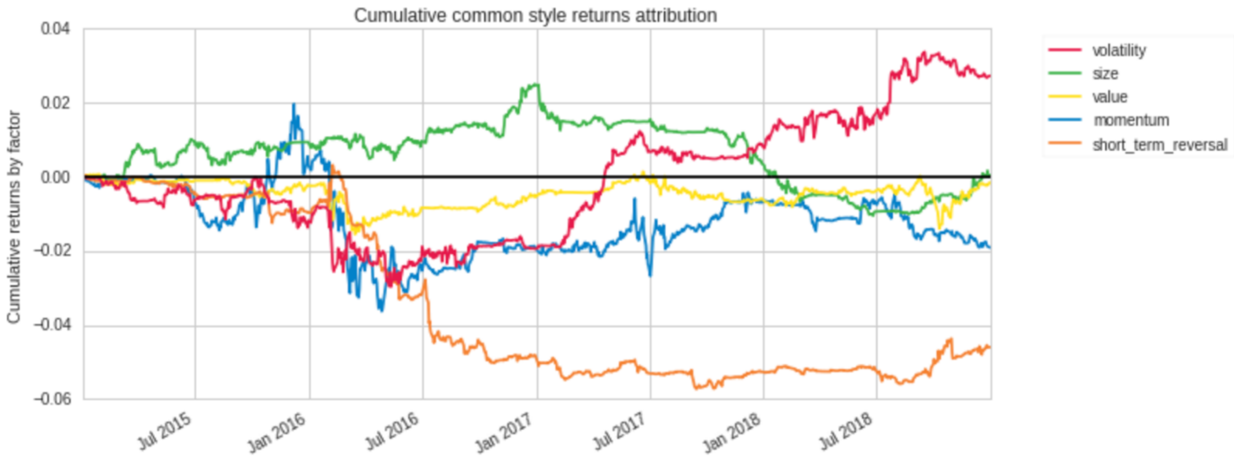

Figure 17 - Cumulative Returns of Style Factors (Strategy 2)

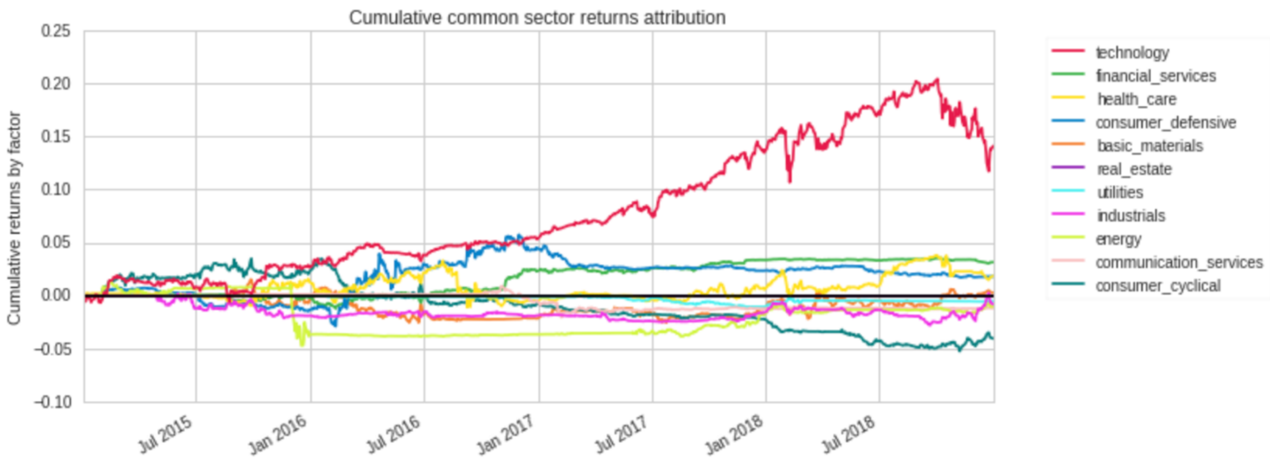

Figure 18 - Cumulative Returns of Sectors (Strategy 2)

Figure 19 - Cumulative, Specific and Common Returns (Strategy 2)

From figure 19 we can understand that the strategy performs well with respective common risk factors which are the 11 sector and 5 style factors shown in table 5 above. These risk factors make up the common returns. The specific returns are the difference between the total returns and common returns.

Figure 20- Cumulative Returns Against Benchmark (Strategy 2)

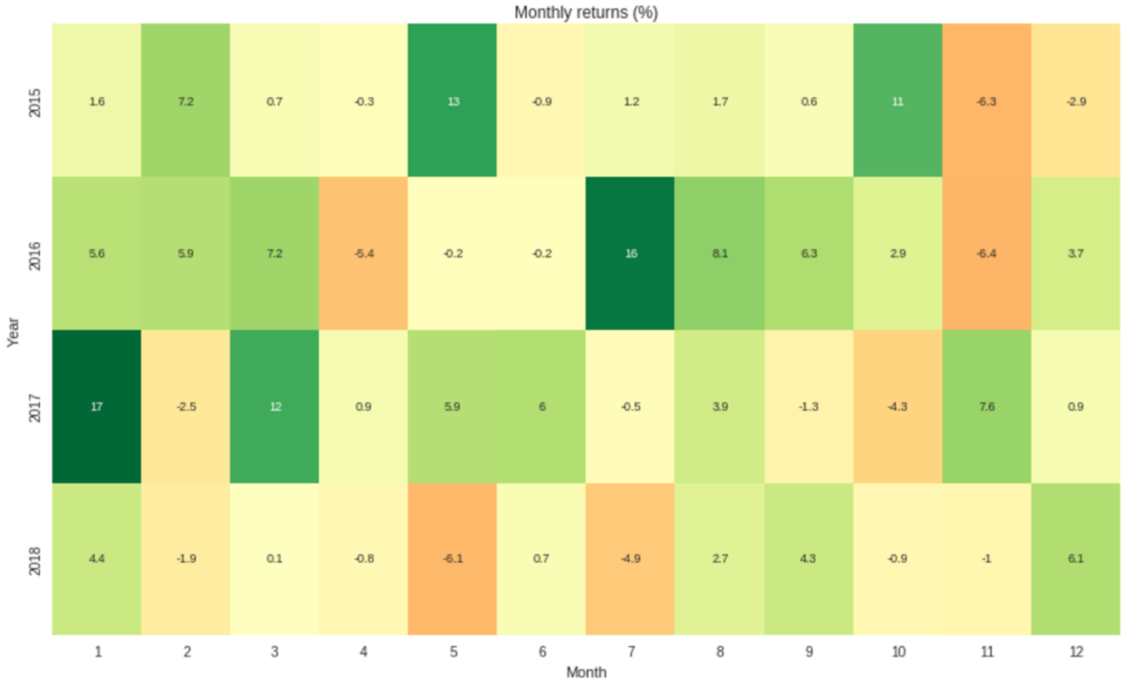

Figure 21 - Heatmap of monthly and annual returns (Strategy 2)

Figure 21 shows the best and worst performing months for the strategy (green to red). January 2017 and July 2016 were top performing months with 17% and 16% monthly returns, while November 2016 and November 2015 were the worst with -6.4% and -6.3% returns respectively.

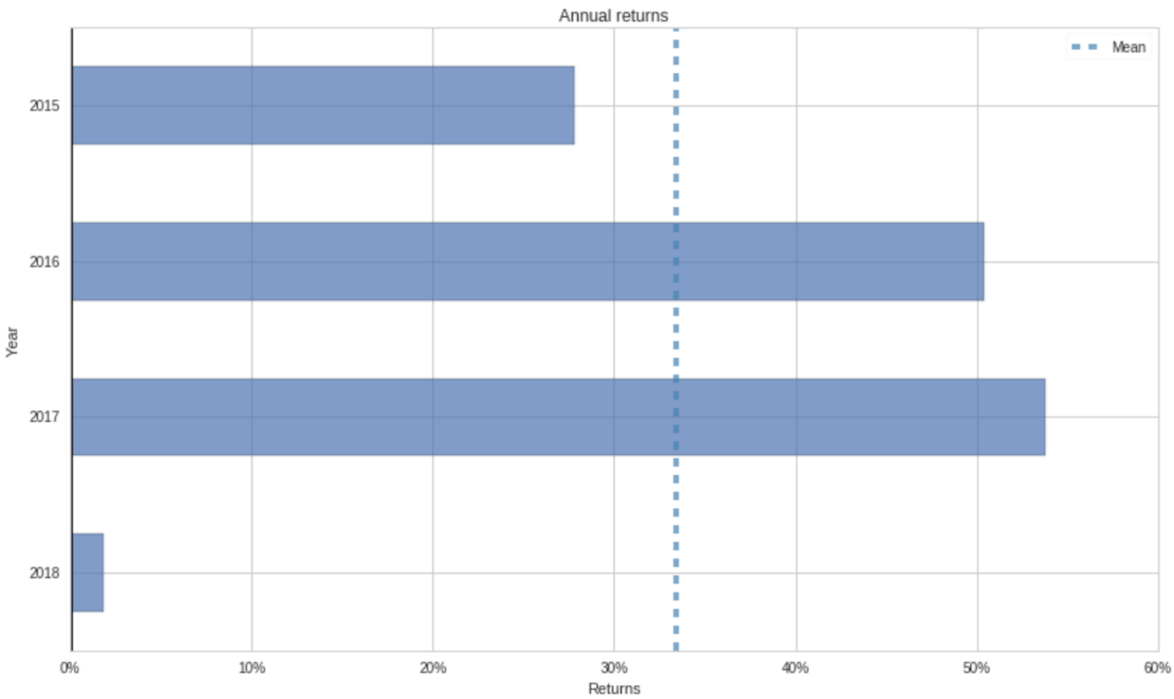

Figure 22 - Annual Returns Histogram (Strategy 2)

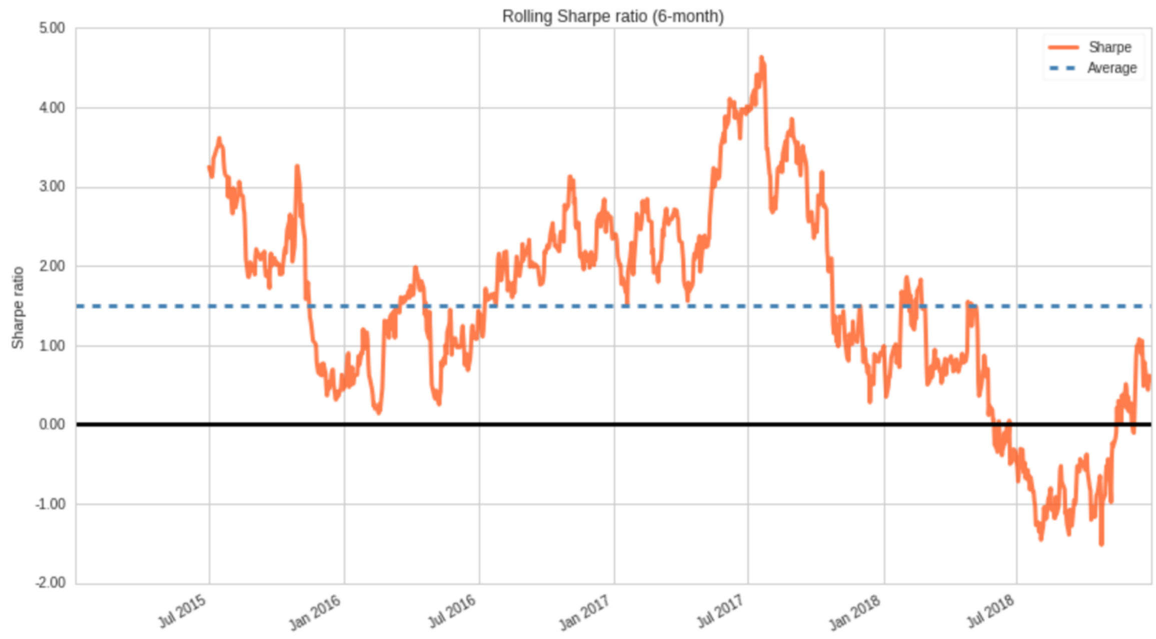

Figure 23 - Rolling Sharpe Ratio (Strategy 2)

The Sharpe ratio sees a sharp decline starting from the end of 2017 to the middle of 2018 in figure 23. This period can be further analyzed looking at the maximum drawdown visualizations in figure 24 and 25. Figure 24 plots the top 10 worst drawdown periods along with the cumulative returns during that period.

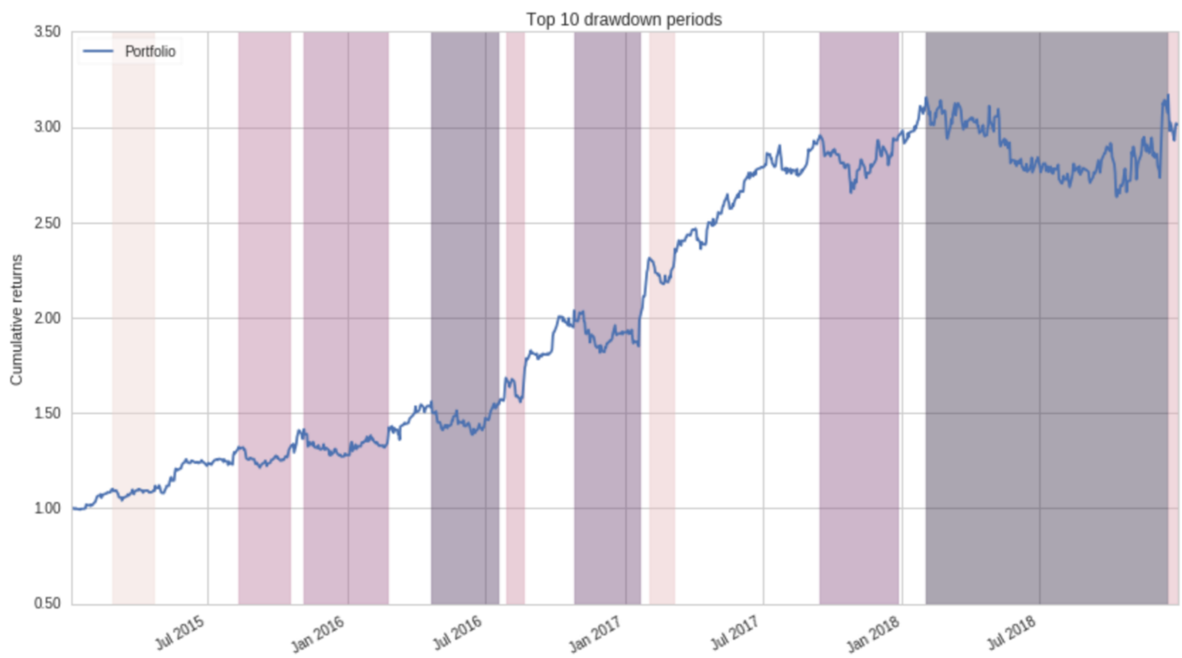

Figure 24 - Top 10 Drawdown periods (Strategy 2)

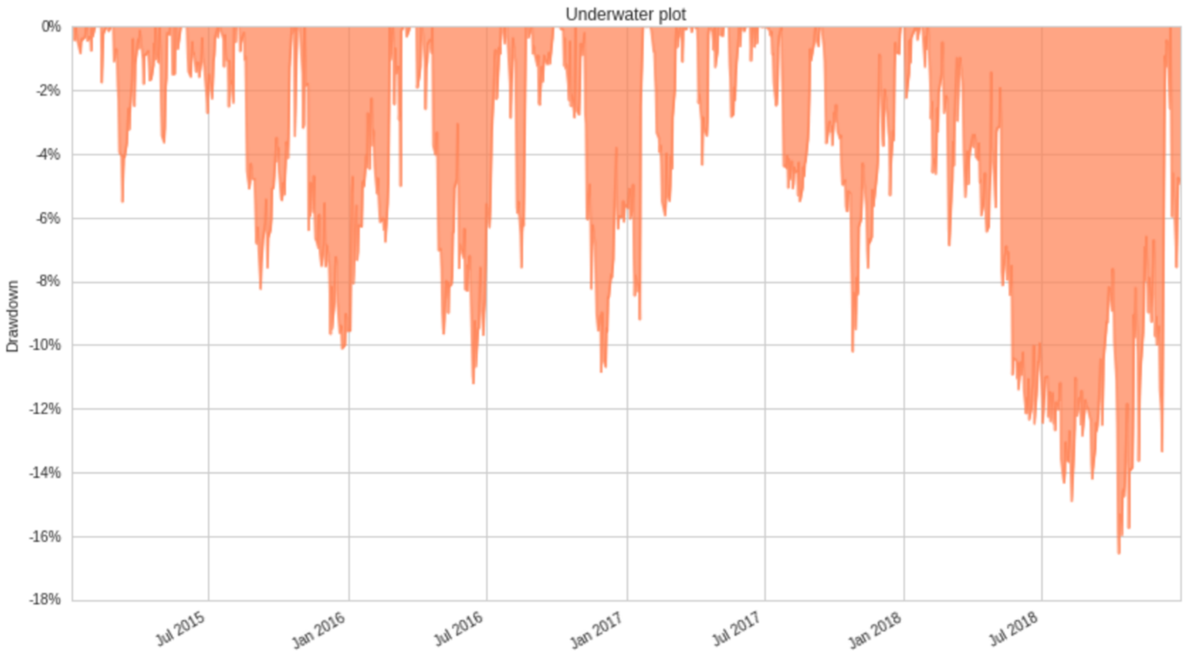

Figure 25 - Underwater Drawdown (Strategy 2)

Figure 25 is an underwater plot to help quantify the drawdowns during the backtesting time period. We can see the longest and highest maximum drawdown occurred in 2018. To get more information about the drawdowns, table 7 contains information about the net drawdown percentage, the duration in number of days, the peak, trough and recovery dates.

Table 7 - Top 5 worst drawdowns (Strategy 2)

Worst drawdown periods	Net drawdown in %	Peak date	Valley date	Recovery date	Duration
0	16.55	2018-02-02	2018-10-11	2018-12-18	228
1	11.21	2016-04-21	2016-06-14	2016-07-20	65
2	10.85	2016-10-26	2016-11-29	2017-01-23	64
3	10.21	2017-09-15	2017-10-26	2017-12-28	75
4	10.12	2015-11-05	2015-12-24	2016-02-25	81

6.3. Comparison of Strategies

Although both the strategies beat the S&P500 benchmark index, they have different results in terms of performance, cumulative total and specific returns, exposures to sectors and factors, volatilities, etc. Table 8 is a comparison of both the strategies performances during the time period 2015-01-01 to 2018-12-31. Figure 26 plots the cumulative returns of both the strategies.

Table 8 - Comparison of Strategy Performances

Performance Metric	Strategy 1	Strategy 2	S&P500 Benchmark
Cumulative Returns	102.12%	201.25%	72.82%
Annual Returns	19.30%	30.85%	11.58%
Annual Volatility	13.31%	20.30%	13.38%
Sharpe Ratio	1.39	1.46	0.88
Maximum Drawdown	-11.78%	-16.55%	-19.30%
Alpha	0.15	0.28	~
Beta	0.40	0.24	~

Looking at table 8, the strategies have different risk-returns ratios. Strategy 1 would seem more suitable for an investor who is looking for an active investment strategy but willing to take same level of risk as compared to the S&P500 which has annualized 4 year returns of 11.58% with annual volatility of 13.38% while experiencing larger drawdown as well.

Strategy 2 is suitable for an investor with a higher appetite for risk. Annual returns of 30.85% which is almost 5x higher than the S&P500 annual returns for a slightly higher risk appetite in comparison.

Figure 26 - Cumulative Returns of Strategy 1 & 2

7. Conclusion and Further Work

The project was successfully completed with both the trading strategies meeting the objectives of beating the S&P 500 benchmark index by implementing different heuristic mathematical optimization techniques such the minimax optimization and the mean-variance optimization along with machine learning models used to diversify portfolios, predict the future change in price and volume.

The results of the strategies show promise of utilizing machine learning techniques in portfolio management. The strategies can be further optimized with extensive fundamental research to build better alpha factors, reduce gross exposures to different underperforming sectors and style factors in order to achieve higher risk-adjusted returns. The strategies can be tested over longer time periods and also during financial crisis periods to evaluate the performances when the market is different. I will be further developing the algorithms and strategies for the Quantopian trading algorithm contest.

Although there is no widely used machine learning strategies for active portfolio management, I hope this project and research study can be utilized in the practice as a beginning step and be further developed and improved. Further research can be conducted on different machine learning approaches and hyperparameter optimization for better results.

The strategies are evaluated using historical data and backtesting to verify how the strategy would have performed. This is only the initial stage of the evaluation process to research and ensure the strategies are fundamentally strong and sound. The strategies need further enhancements before an investor uses them in a live trading environment.

Appendix I:

Both strategies can be found along with the research notebooks on my GitHub Repository link

https://github.com/suhas98?tab=repositories

Bibliography

[1] N. Lioudis, "Passive vs. Active Portfolio Management: What's the Difference?," 15 February 2020. [Online]. Available: https://www.investopedia.com/ask/answers/040315/what-difference-between-passive-and-active-portfolio-management.asp.

[2] T. Segal, "Fundamental Analysis," 16 March 2020. [Online]. Available: https://www.investopedia.com/terms/f/fundamentalanalysis.asp.

[3] "What is factor investing?," BlackRock, [Online]. Available: https://www.blackrock.com/us/individual/investment-ideas/what-is-factor-investing.

[4] J. Chen, "Factor Investing," 16 March 2020. [Online]. Available: https://www.investopedia.com/terms/f/factor-investing.asp.

[5] J. Chen, "Piotroski Score," 22 June 2019. [Online]. Available: https://www.investopedia.com/terms/p/piotroski-score.asp.

[6] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," 08 June 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[7] P. Grover, "Gradient Boosting from scratch," 09 December 2017. [Online]. Available: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

[8] M. Young, "A Minimax Selection Portfolio Rule with Linear Programming Solution," Ann Arbor, 1996.

[9] K. Marvin, "Creating Diversified Portfolios Using Cluster Analysis," 2015.

[10] S. Bülow, "The Effectiveness of Fundamental Analysis on Value Stocks – an Analysis of Piotroski's F-score," 2017.

[11] J.-H. M. Mohr, "Utility of Piotroski F-Score for predicting Growth-Stock Returns," Frankfurt.

[12] C. B. Murphy, "Earnings Before Interest and Taxes – EBIT," 02 April 2020. [Online]. Available: https://www.investopedia.com/terms/e/ebit.asp.

[13] C. B. Murphy, "Enterprise Value – EV," 20 June 2019. [Online]. Available: https://www.investopedia.com/terms/e/enterprisevalue.asp.

[14] J. Maverick, "Using EV/EBITDA and Price-to-Earnings (P/E) Ratios to Assess a Company," 05 February 2020. [Online]. Available: https://www.investopedia.com/ask/answers/072915/how-can-evebitda-be-used-conjunction-pe-ratio.asp.

[15] S. Madge, "Predicting Stock Price Direction using Support Vector Machines," 2015.

[16] T. Yiu, "Understanding Correlation And Diversification," [Online]. Available: https://towardsdatascience.com/understanding-correlation-and-diversification-661c19a26555.