A comparison of a factor-based investment strategy and machine learning for predicting excess returns on the JSE

By

Stefan Drue

Supervised by

Associate Professor Deshen Moodley

A dissertation presented for the degree of

Mphil in Information Technology

Department of Computer Science

University of Cape Town

November 2018

DECLARATION

I, ……………………………………, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: ……………………………………
Signed by candidate

Date: ……………………………………….

Abstract

This study investigated the application of Machine Learning to portfolio selection by comparing the application of a Factor Based Investment strategy to one using a Support Vector Machine performing a classification task. The Factor Based Strategy uses regression in order to identify factors correlated to returns, by regressing excess returns against the factor values using historical data from the JSE. A portfolio-sort method is used to construct portfolios. The machine learning model was trained on historical share data from the Johannesburg Stock Exchange. The model was tasked with classifying whether a share over or under performed relative to the market. Shares were ranked according to probability of over-performance and divided into equally weighted quartiles. The excess return of the top and bottom quartiles was used to calculate portfolio payoff, which is the basis for comparison. The experiments were divided into time periods to assess the consistency of the factors over different market conditions. The time periods were defined as pre-financial crisis, during the financial crisis, post financial crisis and over the full period.

The study was conducted in the context of the Johannesburg Stock Exchange. Historical data was collected for a 15-year period – from May 2003 to May 2018 – on the constituents of the All Share Index (ALSI). A rolling window methodology was used where the training and testing window was shifted with each iteration over the data. This allowed for a larger number of predictions to be made and for a greater period of comparison with the factor-based strategy. Fourteen factors were used individually as the basis for portfolio construction. While combinations of factors into Quality, Value and Liquidity and Leverage categories was used to investigate the effect of additional inputs into the model. Furthermore, experiments using all factors together were performed.

It was found that a single factor FBI can consistently outperform the market, a multi factor FBI also provided consistent excess returns, but the SVM provided consistently larger excess returns with a wide range of factor inputs and beat the FBI in 12 of the 14 different experiments over different time periods.

Table of Contents

TABLE OF FIGURES

Glossary of Terms

Term	**Meaning**
Share/Stock	A unit of ownership in a publicly traded company that represents a portion of the company’s value.
Portfolio	A collection of shares and/or other financial assets held by an individual, which is constructed in accordance with investment objectives.
Portfolio Rebalancing	The act of selling and buying shares in order to return the portfolio to a state of compliance with its objectives.
Stock Market/Exchange	A place where the buying, selling and issuing of stocks/shares of publicly-traded companies occurs.
JSE (Johannesburg Stock Exchange)	The main stock exchange in South Africa
Return	The % change in the price of a share relative to the purchase price of the share.
Excess Return	The return earned by a share or portfolio, less the return of the market over the same period
Financial Ratio	The relative magnitude of two or more items from a company’s financial statements
Investment Horizon (Short or Long Term)	The duration of time that you plan to hold your investment.
Active Investment Strategy	A ‘hands on’ strategy where a portfolio manager actively seeks excess returns by trading in shares over the short term to benefit from short term price fluctuations.

Passive Investment Strategy	A ‘hands off’ strategy where by shares are bought and held for long periods of time. This strategy aims to minimize trading costs that erode returns under an active strategy.
Factor Based Investment Strategy	An investment strategy that bases its investment on companies financial ratios.
Portfolio-sort	A method of selecting a factor based investment portfolio by sorting the shares based on a factors value and taking opposite positions in the top and bottom groups of shares.
Market Efficiency	The ability of the market to incorporate new information into the share price.
Payoff	The sum of the positive return on the top portfolio and negative return on the bottom portfolio.
Share Index	A measurement of a selection of stocks that is used as a tool to benchmark investments.
Average Portfolio Payoff	The average payoff of the series of portfolios formed on a factor, over a period of time.

1 Introduction

At its simplest, the stock market consists of buyers and sellers who trade in the shares (also referred to as stock, or equity) of publicly listed companies. Each share represents an ownership claim on the business. The value of each share is determined by market forces. Where demand for a share is high, the price increases and where demand is low the price decreases. Many different aspects influence the demand for a particular share. These range from behavioral, political, and psychological to the financial performance of the entity. The ultimate goal of a market participant is to earn a return on their investment, by selling their shares for more than what they paid for them. However, simply earning a return is not enough. To fully justify a strategy of actively choosing stocks rather than investing in a passive index tracker, one should aim to achieve returns in excess of the market return.

Companies listed on the JSE release financial statements each year. These financial statements communicate the financial performance of the company over the previous 12 months. Many aspects of a company's financial performance can be inferred through financial ratios. A financial ratio is the relative magnitude of two or more items from a company's financial statements. These ratios allow us to easily get an understanding of common performance metrics by being condensed into comparable values.

There exist many strategies for making investments in the stock market. Movements such as value investing, pioneered by Benjamin Graham and used by Warren Buffet, contrarian investing, momentum trading and many others are active investments strategies where the investor buy and sell the shares they will invest in based on some form of metric or information. There also exist passive strategies, whereby the investor buys into an index or exchange traded fund (ETF), which aims to track the price of the market or a subset of the market. The market is the investable shares listed on the JSE, or a subset of those shares which would be an index such as the FINDI which is an investable asset composed of the 30 largest financial and industrial shares on the JSE. In this strategy, the investment is held for long periods of time without any changes being made to the composition of the investments. This strategy aims to generate better returns by lowering the costs involved with an active strategy, as transaction costs can be a significant expense for managed funds. A factor-based investment strategy aims to build portfolios that provide consistently better returns than the market or share index based on the underlying financial ratios of the firms listed on the stock exchange. This strategy is based on empirical work on stylized facts which are explained in section 2.2.

Machine learning techniques have the potential to find excess returns with better consistency and with greater magnitude than traditional techniques. Many Machine Learning techniques applicable to problems of this nature make use of classification. Classification is the task of using statistical models to identify to which category a new observation belongs. The models are trained using data with known categories. The elements that make up each unit of training data are referred to as features. There are many different approaches to these problems, with a variety of machine learning techniques yielding positive results. Various types of neural networks [1] and support vector machines [2] have shown to be promising.

Studies on the stock market vary in their approach, with some having a short-term horizon and looking at the stock market over periods of hours and days (akin to trading). Other studies have longer investment horizons and view the market over weeks, months or years (akin to investing). Studies use a variety of input features, depending on the nature and objectives of the research. Machine Learning algorithms have been shown to be adept at using market information to make an informed decision about future market events. These can be as diverse as the movement of a market index in the short term, predicting firm bankruptcy, or finding long term investment strategies to outperform the market.

1.1 Research Aims and Objectives

The aim of this study is to compare a factor-based strategy with a machine learning strategy for predicting excess returns on the JSE.

The objectives of this research are:

- Review empirical evidence on stylized facts that underpin a factor-based investment strategy (FBI)
- Implement and evaluate a traditional FBI on the JSE
- Implement and evaluate a Machine Learning strategy, using a Support Vector Machine (SVM) on the JSE

- Compare the performance of the models using suitable evaluation measures

1.2 Tools and Approach

The research approach involved gathering information on stylized facts and factor-based investment strategies as the basis for the research. Applications of machine learning in financial markets were then analysed in order to determine the best approach to take. This revealed SVM’s as the most prudent algorithm to implement due to both performance and complexity. A dataset containing financial data from the JSE was collected. Progressively more complex models were then made in Python, in order to analyse the data and generate results.

Financial data is sourced through Thomson-Reuters Eikon terminals. This data is exported into .csv format to be used for analysis in Python. The code for the study is written in Python (3.6). A number of libraries are used. These include pandas and sci-kit learn. Results are copied to a Microsoft Excel workbook in order to be stored and properly analysed.

1.3 Contributions

The primary contribution of this work is the application and evaluation of machine learning as an investment tool, based on a factor-based strategy. This work contributes to the field of study by comparing a traditional investment strategy against an SVM. This is one of a few studies to take this approach. The research showed that SVM’s can identify excess returns with greater magnitude and consistency than traditional methods.

1.4 Structure of Dissertation

A review of the related work is presented in chapter 2. Chapter 3 provides the experimental design and implementation of the experiments. Results are presented and analysed in Chapter 4. A summary of key findings, a discussion on possible future work and the study’s conclusion are presented in Chapter 5

2 Literature Review

This chapter is a critical assessment of the related work. It begins with an outline of the stock market and the efficient market hypothesis. It then introduces stylized facts on the Johannesburg Stock Exchange. An explanation and summary of the relevant work is conducted, including the assessment of the methodologies used in these studies. Factor based investment strategies are then introduced, including the implementation of such strategies. A brief introduction to Machine Learning is made, followed by an assessment of machine learnings applications in the financial world. The respective machine learning algorithms used are examined within the context of their application and their suitability assessed. The findings of the literature review are then presented.

2.1 The Stock Market

The stock market is a complex system that has been the focus of numerous studies. From a financial perspective, the ultimate goal of an investor would be a model that would predict stocks that would provide greater returns than the market – 'to beat the market' – over the long term.

The Efficient Market hypothesis (EMH) has been a central tenet of finance since the 1970s [3]. It states that all past and present information regarding a stock is captured within the current stock price, this means that no stock is ever mispriced [3]. This implies that an investor cannot consistently beat the market with an active strategy. However, evidence exists that the market is not perfectly efficient. In his research note Sewell [4] reviewed the notable literature relating to the EMH. His chronological review of over 150 papers from the early 1900s, through the formalisation of the EMH in the 1970s, to the early 2000s provides an understanding of the development and important findings with regards to the EMH. Sewell concludes that just under half the papers reviewed support the EMH [4]. He notes that "fully efficient" is an exacting standard that is unlikely to be met. Lower levels of market efficiency imply that inefficiencies exist, which may present exploitable opportunities.

If inefficiencies exist in the market, in the form of mispricing or uncaptured risk, it would allow an investor to exploit these inefficiencies in order to beat the market. There is a wealth of literature on this topic and within the South African market a number of inefficiencies have been identified, that have been persistent over a number of years. These indicate that a strategy may exist to earn excess returns for a given level of risk. These persistent inefficiencies are termed “stylized facts” or succinctly as “styles” in the literature [5-8]. A stylized fact is defined by Sewell as “a term used in economics to refer to empirical findings that are so consistent (for example across a wide range of instruments, markets and time periods) that they are accepted as truth” [9].

2.2 Stylized Facts on the JSE

There are numerous studies on the stylized facts present on the Johannesburg Stock Exchange (JSE). Before continuing it is important to clarify “styles/stylized facts” vs “Factors”. Factors are all the inputs into the model. Stylized facts are the factors that show persistent correlation to excess returns (as defined above). Thus, it can be said that while all styles are factors, not all factors are styles. To demonstrate this point, Dividend Yield (DY) is a factor under investigation in this study. However, in the pre-crisis period it had the third worst p-value of 0.020599052 as seen in Table 4, and worst payoff of 0.610294063% in Table 5. This means it is not a stylized fact as it does not show a consistent relationship with excess returns.

Many of the studies build on the work of Van Rensburg [7] who uses dividend adjusted monthly return data from Industrial shares listed on the JSE between 1983 and 1999. Van Rensburg examined over 20 style strategies using a portfolio-sort based approach and finds 11 statistically significant styles after adjusting for risk. He uses cluster analysis to conclude that three style groups exist: “earnings to price” as the ‘value cluster’, “market capitalisation” as the ‘quality cluster’, and “12 month past positive returns” as the ‘momentum cluster’.

Using the same style strategies as the 2001 study, Van Rensburg and Robertson [8] employed the standard cross-sectional regression procedure from Fama and Macbeth’s 1973 study [10] and a multifactor model. Using the monthly dividend adjusted data from between 1990 to 2000, obtained from the BARRA organization, they found individually: Price-to-NAV (Net Asset Value), Dividend Yield, Size, Price to Earnings and Cash flow to Price to be significant factors. When constructing a multi-factor model however, they find only two significant styles: Size (as log of market capitalization) and Price to Earnings (P/E) - despite price to NAV being the most significant factor on a univariate basis.

Using the same dataset as Van Rensburg and Robertson [8], Auret and Sinclaire examine the five most significant styles from the 2003 study as well as adding the Book to Market Ratio (BTM) [11]. Using multiple regression analysis, they find that when Book to Market is added to the Van Rensburg and Robertson model it almost completely subsumes the effects of size and P/E. However, BTM does not improve upon the Van Rensburg and Robertson model due to correlation with other significant factors. They note that the Size and PE model has explanatory variables with very low correlation which contributes to its success [11].

Hoffman [12] followed the cross-sectional regression and sorted returns method used by Fama and French in their 2008 study of anomalies [13]. Using dividend adjusted returns on the JSE from 1985 – 2010 and controlling for corporate actions and survivor bias and using both an equal weighted and market value weighted approach, Hoffman finds support for size, book to market and momentum effects. As well as to a lesser significance, earnings to book effect and a new-shares in issue effect.

In arguably the most extensive study of the topic, Muller and Ward [6] use twenty seven years of JSE share price data from 1985 to 2011. They exclude companies outside the ALSI - the top 160 shares on the JSE, which accounts for 99% of its market capitalization - unlike previous studies. Using a portfolio sort methodology, they find that 12-month momentum is an important style (which is consistent with prior literature) and that a combination of momentum, return on capital, cash-flow to price and earnings yield give the best persistent outperformance of the JSE. They do not however include Book to Market in this study.

Kruger and Toerien employed a univariate regression based analysis to investigate the predictability of share returns on the JSE during the period 2000-2009 [5]. They note that previous studies did not distinguish the state of the market over the analysis period. Part of their study aimed to determine if these stylized facts persist during times of market instability. They split the data from 2000-2007, for the bull market period and from 2007-2009 for the period of market instability due to the financial crisis. During the financially stable period they found eight significant style factors that aligned with those identified in previous work. However, during the period of market instability, none of these previously identified style factors persisted and only cash-flow to price was significant. They conclude that the evidence suggests persistent stylized facts on the JSE do exist.

From the literature, it is evident that the JSE is not a fully efficient market, as persistent stylized facts are found over multiple periods. These stylized facts demonstrate a means of identifying shares that will outperform the market and imply a factor-based investment strategy may be able to consistently beat the market. A summary of findings is presented below, in Table 1.

TABLE 1: STYLIZED FACTS - SUMMARY OF FINDINGS

Study	Significant Styles Identified	Methodology
Van Rensburg [7](2001) Years of Data = 16	• Earnings to Price • Market Capitalisation • 12 Month Past Positive Returns (momentum)	• Portfolio sort
Van Rensburg and Robertson [8] (2003) Years of Data = 10	• Size • Price to Earnings	• OLS Regression
Auret and Sinclaire [11] (2006) Years of Data = 10	• Book to Market	• Univariate and Bivariate Regression
Hoffman [12] (2012) Years of Data = 25	• Size • Book to Market • Momentum • Earnings to book • New shares in issue	• Cross sectional regression and portfolio sort
Muller and Ward [6] (2013) Years of Data = 27	• 12 Month Momentum • Return on Capital • Cash-flow to price • Earnings Yield	• Portfolio Sort using a VBA engine
Kruger and Toerien [5] (2014) Years of Data = 9	• Book to Market • Cash-flow to Price • Earnings Yield • Size • 6-month Momentum • 12-month momentum	• Ordinary Least Squared Univariate Cross-sectional Regression

2.3 A Factor Based Investment Strategy

A factor-based investment strategy uses financial ratios from the firms listed on the stock exchange as the basis for its investment decision. The underlying logic of this investment strategy is that share performance can be linked to a firm's financial ratios. It supposes that firms with good ratios perform well and firms with bad ratios perform poorly. In order to identify which financial ratios have the strongest relationship with returns a regression analysis is performed. Once ratios with a statistically significant relationship are found, share portfolios can be formed. These portfolios are formed based on a portfolio-sort methodology.

Regression is a statistical tool used to measure the correlation between one variable and the corresponding values of other variables [14]. It is used to model the relationship between the dependent variable, y, and the explanatory variable or variables, X. Linear Regression can be done with both time series data (Data that is regularly measured over time in a sequential order) or cross-sectional data (Data that is observed or measured at a point in time) [14].

Regression that is performed between a dependent and a single explanatory variable is known as univariate regression. When multiple explanatory variables are used, this is known as multivariate regression [14]. Many estimation methods exist for linear regression, one of the most widely used being Ordinary Least Squares (OLS). This method aims to find the best estimate by minimizing the squared error between the observed dependent variable and the predicted linear function [14].

In the studies summarized in Table 1, two methods are used for identifying or implementing a factor-based strategy, i.e. regression and portfolio-sort. Regression was used to establish whether certain characteristics are able to explain or predict realized share returns [10]. This is done by regressing the factor value against excess returns for each week or month of data and building a time series of regression coefficients. The time series of regression coefficients is then tested for significant difference from a mean of 0. This is provided by the p-value from a student's t-test [15].

The second methodology used is known as portfolio-sort (or sorted returns). Shares are sorted on a regular time interval into fractiles (portfolios) based on one or more firm characteristics [15]. For example, at a point in time, shares are sorted based on a factor, the top and bottom portfolios are created from a subset of the share population. The difference in returns between the top and bottom portfolio represents the payoff of the feature in question. The greater the payoff, the more a feature is seen as significant [15]. This methodology provides a more tangible way to understand the results of the experiments as the payoff of different factors can be compared, and it keeps the results in the metric that is most important to investors. i.e. returns.

When implementing a FBI strategy an investor uses the regression results to identify the most significant factors and uses these factors as the basis of their investment strategy. The portfolios formed by the investor are based on the portfolio-sort method. It can also be used as a justification of the regression results, where the regression informs us of the most significant features and the portfolio sort confirms it by providing the payoff of that feature.

2.4 Machine Learning

Machine learning can be defined as the field of study that gives computers the ability to learn without being specifically programmed [16]. In practice it involves using advanced statistical methods and large quantities of data to make predictions on, classify, or cluster the data. Each machine learning experiment consists of broadly the same set of steps. These are: collecting the data, preparing the data, training the model, testing the model and evaluating the results.

The data set for each machine learning experiment is dependent on the task that one wishes to perform. For a supervised machine learning task, the data can be financial data, images, sounds or any form of data that can be distinctly categorized. Each observation in the dataset will be labelled as belonging to a class. The classes represent the category the information belongs to. The objective of the machine learning model is to train on part of the data and build a relational model that can then be extended to unseen data in order to correctly classify the unseen data into a class based on the input variables. Once the data is collected, it is often necessary to clean and prepare the data. This step is data pre-processing.

Data pre-processing involves removing data points with missing values or outliers and ensuring the validity, integrity and consistency of the data. Often it is necessary to normalise the data. Three main types of normalization can be performed. These are: re-scaling, standardization, and scaling to unit length. Rescaling scales values into a range of either [0,1] or [-1,1] using minimum and maximum values. Standardization normalizes a factor around a mean of 0 and a standard deviation of 1. Scaling to unit length divides each item of the feature vector by the feature's Euclidean length. This process helps algorithms when dealing with features with different ranges, such as a dataset with house prices in millions, area in thousands of meters squared, and room numbers in single digits. It prevents a larger feature from overwhelming a smaller feature.

A further step in data processing is feature extraction. This is the selection of the features that will be used to train the model. In this study, features will include volume, return on equity, book to market ratio and a number of other financial ratios and share data. These ratios are based on the findings of the literature. This process also involves removing any features that do not contribute to the model, as including these may cause error and bias that reduces the accuracy of the model. It can also involve the creation of new features by combining different features; for example, in a housing dataset, creating a new feature by dividing floor size by number of rooms to get an average room size feature – more complex operations can also be performed.

Once the data has been prepared, the data set is split into training and testing data sets. Often the data is split 70/30 training/testing, this is called the hold out method. However, if there is not a large enough dataset or if the nature of the dataset allows for it, cross validation may also be used. When cross validation is employed the data is partitioned into *k* fractiles, then a portion of the fractiles are used as a training set and the rest are used as a test set. This allows for more training to be done as the fractiles can be mixed until all combinations of the data have been used. The results can then be aggregated and averaged. This is necessary as there will be *k* sets of results, each for a different partition of the data. Therefore, the average of the results is taken to get an average result for the model.

A common issue is overfitting the data. This occurs when the algorithm incorporates too much noise and detail into its solution on the training data. This prevents it from generalising well to the test data or new data. Cross validation is often an important step to diagnose overfitting before the test data is analysed.

Once an algorithm has been tested, the results need to be evaluated for accuracy. Depending on the nature of the task, different measures may be utilised. In classification tasks popular metrics include classification accuracy (or hit ratio in some studies) which is the ratio of correct classifications to all predictions made. Other metrics such as logarithmic loss, which gives a scalar probability between 0 and 1 that can be seen as a measure of confidence in the prediction, and area under the ROC curve, which measures a model's ability to discriminate between positive and negative classes as its discrimination threshold is varied. A popular evaluation tool is the confusion matrix. The confusion matrix allows for the visualization of the model's performance. It sets out the actual classes of the test against the predicted classes and displays both the classification accuracy in determining true positives and negatives as well as the false positives and negatives. This information is more important than the classification accuracy, particularly with unbalanced datasets. The confusion matrix provides a large amount of useful information in a compact, easy to read structure.

TABLE 2: CONFUSION MATRIX DIAGRAM

Confusion Matrix		Actual Class	
		Class A	Not Class A
Predicted Class	Class A	True Positive	False Positive
	Not Class A	False Negatives	True Negatives

Another useful evaluation metric is precision, recall and F1 score. Precision shows how precise or accurate a model is at determining the true positives. It is calculated as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Precision is valuable when the cost of a false positive is high [17]. Conversely, when the cost of a false negative is high, Recall is valuable as it shows the ratio of correctly predicted observations to all the actual observations in the positive class [17]. Recall is calculated as follows:

$$Recall = \frac{True\ Positive}{True\ positive + False\ Negative}$$

Using Precision and Recall an F1 score can be calculated as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The F1 score is used when a balance of precision and recall is needed. It is particularly important when there is an uneven class distribution [17].

In regression tasks, popular evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2. MAE measures the absolute error between the predictions and actual values. This provides the magnitude of error but does not give an indication of the direction of error. MSE is similar to MAE, however it uses the square of the error term to provide magnitude of error. R^2 , otherwise known as the coefficient of determination, provides an indication of the goodness of fit of the predictions to the actual values with a score between 0 (poor fit) to 1 (perfect fit).

2.5 Machine Learning in Finance

Machine Learning has emerged as a relevant field due to its ability to predict or classify outcomes given large sets of complex, correlated information [2]. It has seen traditional use in classification and pattern recognition problems. Its malleability has led to it being used in many different applications in the financial world [9].

Wilson and Sharda [18] compared neural networks and multiple discriminant analysis to predict firm bankruptcy, and determined that neural networks performed best [18]. Min and Lee [19] used SVMs, multiple discriminant analysis, logistic regression and a three-layer fully connected back propagation neural network. SVMs were shown to be the best performing algorithm for predicting firm bankruptcy[19].

Machine Learning Algorithms have also found application in predicting a company's credit rating. Lee used an SVM and a number of financial ratios to predict credit ratings [20]. Huang et al [21] applied neural networks to the same problem and also achieved promising results with between 75% - 80% accuracy [21].

Machine learning has also been applied to the task of stock market prediction. Given the many different techniques machine learning provides, studies have found different ways of approaching this topic. Some have used sentiment analysis of social networks to predict stock index movement [22], while many others use more traditional approaches, which are more applicable to this study.

The most relevant works for this study have focused on predicting or classifying the future movement of a share or share index. Two distinct methodologies have been identified for this application – where machine learning is applied to predict the short-term movement of a share index and where it is applied with a longer horizon and predicting or classifying market outperformance.

2.6 Machine Learning for Short Term Index Prediction

The studies identified can be broadly grouped into two distinct categories, those that take a short-term view of the market, and those that take a long-term view of the market. Those that take a short-term view of the market have generally aimed to classify whether a market index will move up or down in the next time period. When viewing the market with a longer horizon, the aim moves away from predicting an up or down movement and instead focuses on finding ways to outperform the market irrespective of its movement. While Machine Learning algorithms provide different ways with which to approach this task, there are common elements that appear throughout the literature - where excess returns are identified based on some fundamental firm data or technical indicators or both.

Cao and Tay [23] use data from the S&P 500 Daily Index in the Chicago Mercantile Exchange, with training data from 01 April 1993 to the end of December 1994 and test data from 01 March 1995 to the end of December 1995. They compare the use of Support Vector Machines (SVMs) to a multi-layer perceptron trained by Back Propagation (BP) for predicting the movement of the index for the following day. They treat this as a classification problem by classifying the predicted direction of movement of the S&P500. They found that SVMs offer a number of advantages for time series forecasting. These include a smaller number of free parameters compared to BP, faster training than BP and better forecasting when measured by Normalised Mean Squared Error (NMSE), Mean Absolute Error (MAE), Directional Symmetry (DS), Correct Up Trend (CP) and Correct Down Trend (CD).

Kim [1] applied SVMs to predicting the stock price index and compared it to BP neural networks and case-based reasoning (CBR). Using data from the Korea composite stock price index (KOSPI) from January 1989 to December 1998 (2928 trading days) Kim attempted to predict the direction of the daily change of the KOSPI price based on 12 technical indicators. The data was split 80% training, 20% testing. He found that SVMs outperformed BP but not by a statistically significant margin, while SVMs significantly outperformed CBR at the 5%

level. He attributes SVMs performance to the implementation of the structural risk minimization principle which leads to better generalisation than traditional techniques.

Inspired by Kim's work, Qian and Gao [24] make a comparison between traditional time series models and machine learning methods. They compare Autoregressive Integrated Moving Average (ARIMA) to Logistic Regression, SVM, Multi-Layer Perceptron (MLP) and a Denoising Autoencoder (DAE) SVM. Using the S&P 500, Dow 30 and Nasdaq indices as their predicting data between January 2012 and December 2016. Their time period was weekly data and they obtained the open price, close price, high and low price as well as trading volume for the day. For the machine learning models, they also included a number of technical indicators and split their data 70/30 for training/testing. The target variable in this experiment was the closing price movement direction for the next trading day. They use a hit ratio to measure precision and found, like Kim (2003), that SVMs are adept at price movement prediction. DAE-SVM and SVMs were the two best performing models with hit ratios of 69.1% and 64.2% respectively, outperforming ARIMA with a hit ratio of 59.3%. Logistic regression managed a hit ratio of 62.3% while MLP performed the worst with a hit ratio of 56.6%.

2.7 Machine Learning for Long Term Market Outperformance

Milosevic [25] treats this as a classification problem. Her model aims to classify whether a share will have a 10% higher price after a 1-year period. Using data from multiple markets, 1739 stocks from the S&P 1000, FTSE 100 and S&P Europe 350, and 28 fundamental firm indicators. She found a Random Forest algorithm produced the best results with 75.1% Precision. She also identified the 11 features that gave the best performance: Book Value, Market Cap, Dividend yield, Best EPS, PE ratio, Price to Book Ratio, Best DPS (Dividend per share), Current Ratio, Quick Ratio, Debt to Equity Ratio and historic price. Using only these features the Random Forest improved with a precision of 76.5%. Her findings support the work of previous literature as her significant features corroborate the stylized facts identified in previous literature.

Fan and Palaniswami [26] also treat this as a classification/pattern recognition problem but use SVMs. They used a number of financial ratios and group these as – Return on capital, Profitability, Leverage, Return on Investment, Investment, Growth, Short term liquidity and

Risk. Using data from the Australian Stock Exchange from 1992 – 2000, they attempted to use the SVM to identify stocks that are likely to produce excess returns – ‘beat the market’ – by creating 4 equally weighted portfolios of the stocks. They found that SVM-created portfolios beat the market in each year of their study. They produced a return of 208%, while the market managed 71% over the same 5-year period. This is also a demonstration that stylized facts exist internationally.

Arik, Eryilmaz and Goldberg [27] note that while machine learning has seen uptake in high frequency trading and market valuation based on economic parameters, a large portion of the finance industry is focused on mid to long term portfolio construction. This is a task still largely done by people. They collected data on 1012 stocks listed on the NYSE over a 10-year period (2004-2013). They selected 69 fundamental financial parameters, which was reduced to 52 parameters after data pre-processing. They constructed their experiment as a classification task and initially considered Decision Trees, K-nearest neighbours, Naïve Bayes and SVM techniques. They settled on an SVM, noting that it is well suited for financial applications due to its capability to model high dimensional feature spaces. Their results showed that the SVM could classify bullish stocks (those that outperform the benchmark – the NYSE composite index) with a 71.2% accuracy in training and 58.8% accuracy in prediction. It classified bearish stocks (those that underperform the benchmark) with 60.02% accuracy in training and 57.9% accuracy in prediction.

In a study that includes many of the important points identified in the literature thus far, as well as being applied to the South African market – making it particularly relevant – Runhaar [2] compares a factor based investment strategy against an SVM used in a classification task. The factors he includes are derived from the literature on stylized facts on the JSE. He selects 100 companies from the JSE and collected 15 years of data from March 2001 to February 2016.

The factor-based investment strategy was an implementation of portfolio-sort, where stocks were ranked based on the factor being tested and then sorted in to quintiles based on their ranking. The first quintile was used as the top performing quintile for comparison with the SVM. For the SVM 10 years of this data was used for training and cross validation, while 5 was used for testing and portfolio performance comparison. He trains the model on the historical factor data and uses this to predict the probability of future upward stock price movement. Stocks were ranked based on probability of belonging to a class and divided into equally weighted quintiles. The first quintiles for both strategies were used as the basis for comparison. He found the Machine Learning outperformed the benchmark for 6 of the 7 factors, outperforming the accuracy of the factor-based strategy significantly. The best performance was attained with the machine learning model using all 7 factors as inputs. The factors used in his study are: Dividend Yield, Earnings Yield, Book to Market Value, 6-month momentum, 12-month momentum, return on equity and return on invested capital. He notes the limitations in his study not accounting for transaction costs and only using a single machine learning algorithm, although prior literature indicates that the best results are often achieved with the use of SVMs.

Runhaar focuses on portfolio construction and constructs portfolios with a 3-month holding period, staggered 1 month apart. I.e.: Portfolio 1 is constructed at the start on month 1 and updated or rebalanced at the end of month 3. However, this results in two untraded months. His solution was to stagger portfolios such that each month a portfolio is created and then held for 3 months before being rebalanced, this allows for a higher frequency of portfolio construction and rebalancing, provides more data for training and testing and does not impact the buy-hold timeframe.

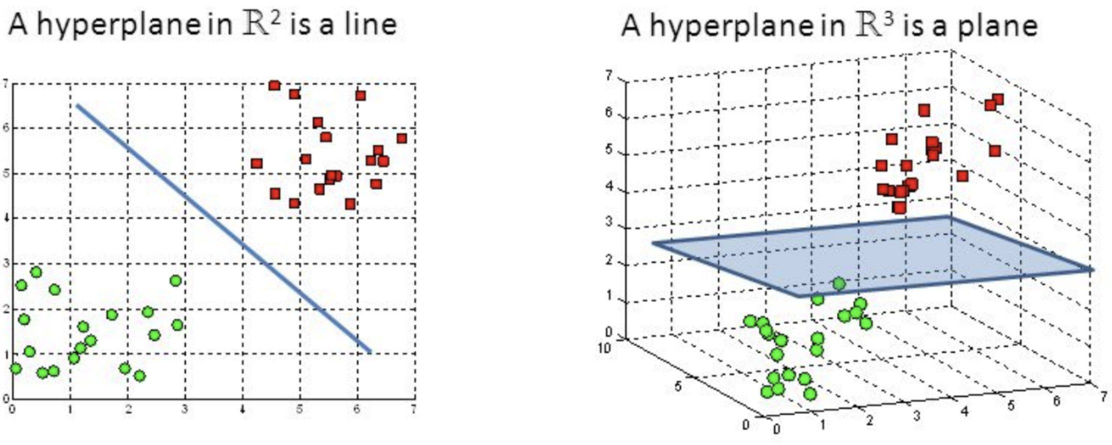
For example: Portfolio 1 (P^1) is created at time t, P^1_t is rebalanced at $t+3$ (P^1_{t+3}). Portfolio 2 is created at $t+1$ (P^2_{t+1}) and rebalanced at time $t+4$ (P^2_{t+4}). Portfolio 3 is created at time $t+2$ (P^3_{t+2}) and is rebalanced at $t+5$ (P^3_{t+5}). The time step in each case is 1 month. Portfolio rebalancing is the process of buying and selling assets in the portfolio in order for the portfolio to comply with the portfolio rules. In practice for a portfolio based on ROE, if a company in your portfolio sees a reduction in ROE such that it no longer is in the top quintile when ranked, you sell your position in that share and buy the share that replaces it. It also is done to ensure the weighting of the portfolio doesn’t deviate too far from the specified weighting.

2.8 Support Vector Machines

The Support Vector Machine (SVM) was a popular algorithm in the reviewed literature and the chosen algorithm for the experiments. The SVM is a classification algorithm introduced by Cortes and Vapnik [28].

The SVM attempts to find the optimal hyperplane in an n-dimensional space, where n is the number of features, such that the hyperplane has the greatest margin between the support vectors of the two classes of data.

The support vectors are the data points from both classes that lie closest to the hyperplane and are the ones that have the greatest effect on the hyperplane. In a 2-dimensional dataset the hyperplane would be a line separating the two classes of data. In a 3-dimensional dataset, the hyperplane is a 2-dimensional plane. This becomes more complex as more dimensions are added. In all cases, the SVMs objective is to find the hyperplane that produces the greatest distance i.e. separability, between the two classes of data. After training, an input vector will result in a point on a side of the hyperplane and the classifier will assign it to one of the classes. The distance of the point to the hyperplane is proportional to the probability of an accurate classification into that class [2].


A hyperplane in $\mathbb{R}^2$ is a line

A hyperplane in $\mathbb{R}^3$ is a plane

FIGURE 1: DEPICTION OF A HYPERPLANE [29]

The SVM classification function is as follows:

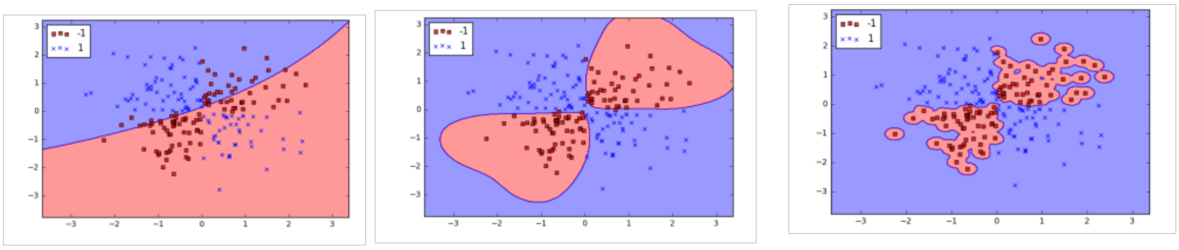
$$f(\mathrm{x}) = \sum_{i=1}^{N} \alpha_i \mathrm{y}_i \mathrm{K}(\mathrm{x}, \mathrm{x_i}) - \mathrm{b}$$

where:

- x is the feature vector consisting of M input variables.
- xi is all historical feature vectors of the training set.
- N is the number of samples used to train the SVM and fit parameters.
- α_i is a scalar tuning parameter with a value between 0 and C.
- y_i accompanies xi by indicating the classification into either the positive or negative class.
- b is a scalar value obtained in the training process that is used to shift the output of the SVM function by a constant.
- $K(x, xi)$ is the kernel function of the SVM. The kernel function transforms the non-separable linear data to a separable higher dimensional space and is critical for the functioning of the SVM.
  Below are the formulae of some popular kernels:
  - $K(x, xi) = \langle x, x' \rangle$ – Linear Kernel
  - $K(x, xi) = (\gamma \langle x, x' \rangle + r)^d$ – Polynomial Kernel
  - $K(x, xi) = exp\,(-\gamma |x - x'|^2)$ – RBF/Gaussian Kernel
  - $K(x, xi) = \tanh\big(\gamma \langle \mathrm{x}, \mathrm{x}' \rangle + \mathrm{r}\big)$ – Sigmoid Kernel

Each kernel has the ability to be tuned through certain hyper parameters. In a RBF/gaussian kernel, the hyperparameters are C and gamma (γ). Each of these parameters has an effect on the decisions boundary of the SVM as depicted in Figure 2 and Figure 3 below.

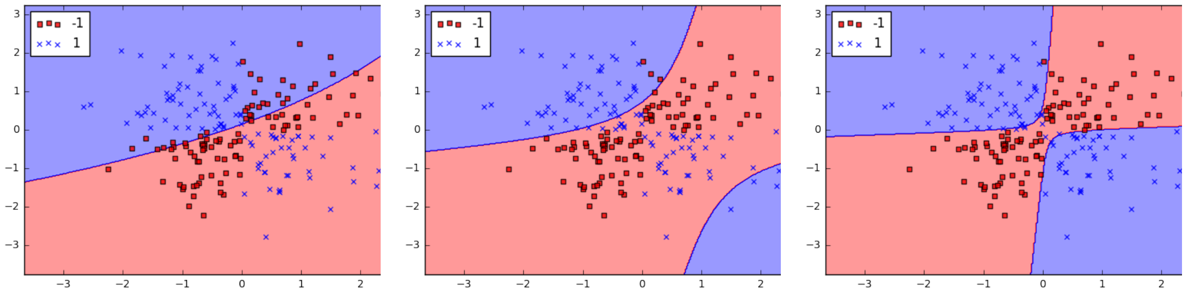
Gamma is the influence a single training example on the decision boundary. Low values of gamma give training examples a large area of influence. As gamma increases, the influence of a training example decreases. Increasing gamma has the effect of making the decision boundary ‘more bumpy’. Too high a value of gamma can result in overfitting and poor generalizability of the model. However, when gamma is too low the model will struggle to capture the complexity of the decision boundary shape. This will also have a negative effect on the classification accuracy of the final model.



$\gamma=0.01$ $\gamma=1$ $\gamma=100$

Figure 2: Effect of different Gamma on decision boundaries [30]

C is the penalty parameter of the error term. It is responsible for controlling the size of the margin in a non-linear SVM. Small values of C will have a larger margin. As the value of C increases, the margin will decrease. Figure 3 demonstrates how increasing C changes the shape of the decision boundary. As C increases the distance between the support vectors of the two classes is decreased. Too large a value of C can result in overfitting. C can intuitively be thought of as controlling the curvature of the decision boundary. The higher the value of C the more curvature the decision boundary will have, but the narrower the margin between the two classes becomes.



C=1 C=10 C=1000

Figure 3: Effect of different C on decision boundaries [30]

2.9 Summary

The literature shows that machine learning techniques can be effectively applied to financial forecasting in a number of manners, in both the long and short term. Machine learning has wide ranging applications from credit risk analysis to predicting stock or market returns or movements. SVMs seemed to be particularly efficient at these tasks. When focusing on long term forecasting, there is evidence that there are exploitable stylized facts on the JSE and that machine learning techniques are adept at identifying and exploiting these.

The literature demonstrates that the South African market has certain exploitable inefficiencies. In studies covering a nearly 30-year period, multiple factors were identified that explained excess returns on the market. These stylized facts, as they are referred to, are identified using multiple methodologies and over different market periods. This demonstrates their persistence and significance and provides a basis from which this can be extended to machine learning.

Machine Learning has shown promise in its application to financial problems. It has been demonstrated usefulness in a myriad of topics and also shows how similar topics can be approached with different methodologies. The literature shows that the application of machine learning algorithms for financial investment yields positive results, often better than what can be revealed through traditional techniques such as factor-based strategies. In particular, support vector machines have been shown to be particularly adept in these tasks. The studies applying the theory behind stylized facts to machine learning [2, 25, 26], have demonstrated improved results versus the traditional methods employed in the literature.

The literature suggests that SVMs provide the most robust and accurate results. SVMs consistently ranked as the best or one of the best performing algorithms in both short- and long-term focus studies. Most studies select a Gaussian kernel, although Linear and Polynomial Kernels were also used. Kim notes that the Gaussian Kernel took less time to train and provides better results [1]. Other popular algorithms included multi-layer perceptron and Neural Networks with back-propagation, Random Forests and Logistic Regression. A training/testing split of between 70/30-80/20 was used for most experiments. Common evaluation metrics included classification accuracy (hit ratio), Precision, recall and F-Score, while for regression-based techniques MSE, MAE and R^2 were often used. Furthermore, comparison between the algorithm and some benchmark return figure (such as market return over the period) are commonly used for stock prediction tasks.

3 Experimental Design

There are many ways to approach the experimental design for any given problem. This is demonstrated in the literature, both by studies on stylized facts and machine learning implementations of investment strategies. In this chapter the specific methodology employed in these experiments is presented. The shared components of both experiments are discussed first. Then Experiment 1 – A factor-based investment strategy - and Experiment 2 – Machine Learning are detailed.

3.1 The Dataset

The dataset consists of financial data for companies on the JSE All Share Index (ALSI) over a 15-year period from May 2003 to May 2018. The dataset is comprised of price data, as well as company specific financial ratios and metrics for the constituents of the ALSI during the time frame. Table 3 shows the different attributes of the data set.

The data was sourced from a Thompson Reuters Eikon terminal at the University of Cape Town. The number of companies varied during the study period to reflect companies that listed or delisted on the JSE. Companies which delisted or were removed from the ALSI were excluded from the data set. The final data set contained data for 97 companies at the start of the study period (May 2003) and increased to 160 companies at the end (May 2018). The dataset is split into four periods for the experiments:

- The pre-crisis period (2003-2007)
- The crisis period (2007-2009)
- The post crisis period (2009-2017)
- The entire period (2003-2017)

TABLE 3: DESCRIPTION OF THE DATASET

Attribute	Explanation	Category	Update Period
Date	The date on which the closing share price is recorded		Weekly
Price	The adjusted weekly closing price		Weekly
Market Capitalization (MC)	The closing price of the share multiplied by the number of shares outstanding	Size	Weekly
Dividend Yield (DY)	Dividend per Share / Price per share	Value	Weekly
PE Ratio (PE)	Price per Share / Earnings per Share	Value	Weekly
Quick Ratio (QR)	Cash and equivalents + Marketable Securities + Accounts Receivable / Current Liabilities	Liquidity and Leverage	Annually
Return on Equity (ROE)	Net Income / Shareholders Equity	Quality	Annually
Return on Invested Capital (ROIC)	Net Income - Dividends / Total Capital	Quality	Annually
Total Debt % to Common Equity (DE)	Total Liabilities / Shareholders Equity	Liquidity and Leverage	Annually
Total Debt % to Total Capital (DC)	Total Liabilities / Total Equity	Liquidity and Leverage	Annually
Price to Book Ratio (PTB)	Share Price / Book value per share	Value	Weekly
Price to Cashflow Ratio (PTCF)	Share Price / Cashflow per share	Quality	Weekly

Return on Assets (ROA)	Net Income / Total Assets	Quality	Annually
Earnings Yield (EY)	Earnings Per Share / Price per Share	Value	Weekly
Market to Book Value (MVTB)	Market Capitalization / Total Book Value	Value	Weekly
Current Ratio (CR)	Current Assets / Current Liabilities	Liquidity and Leverage	Annually
Excess Return	The return earned by a share or portfolio, less the return of the market over the same period	Target Variable	
Payoff	The absolute value of the excess return of the top portfolio – bottom portfolio. Is the equivalent to being long the top portfolio and short the bottom portfolio.	Evaluation Metric	

The target variable for the experiments is excess returns. The excess return is the return earned minus the market return. This is calculated as the change in share price (closing price) over a given period less the change in the average market price over that period. The formulae for returns are given below:

1. $Return = \frac{P_{t1} - P_{t0}}{P_{t0}}$ (for a share)
2. $Market\ Return = \frac{\Sigma\ \mathrm{Return}}{number\ of\ shares}$
3. $Excess\ Return = Return - Market\ Return$

The average market price is the equally weighted return of all the investable shares for that week. The holding period is the length of time that the portfolio is held before selling or rebalancing. The holding period will be 12 weeks and will remain constant throughout the experiments. This holding period was selected as it allows for a good balance between a time period long enough to show the relationship between factors and share price, while providing enough time to create many portfolios required to test the hypothesis.

3.2 Data Pre-Processing

Data preprocessing is an important step in machine learning and regression experiments. Features can exist within large value ranges. Normalizing these ranges helps to create a more balanced dataset and prevents large value input factors from overwhelming small value input attributes. This helps reduce prediction error [1].

The three main types of normalization are rescaling, standardization and scaling to unit length. Rescaling scales values into a range of either [0,1] or [-1,1] using minimum and maximum values. Standardization normalizes a factor around a mean of 0 and unity standard deviation. Scaling to unit length divides each item of the feature vector by the features Euclidean length [2].

In this experiment, data will be standardized across the entire population of the dataset for that specific feature. All data will be standardized around a mean of 0, with unit standard deviation [5]

3.3 Bias, Limitations and Scope

3.3.1 Scope

This study is confined to the South African market. Specifically, to shares listed on the JSE. While all listed shares are within the scope of this study, limitations on the duration and completeness of data and data availability will limit the number of shares included in this study. Further limitations include not accounting for transaction costs in constructing and rebalancing the portfolios. These costs include brokerage fees, Securities Transfer Tax (STT), STRATE fees, FSB Investor Protection Levy and VAT.

3.3.2 Biases

Two significant biases are present in a study of this nature. The first being survivorship bias which occurs when only companies that exist for the entire duration of the study are included.

This ignores the companies that fail during the period and thus can skew the results. Due to the nature of the data collection process, some shares are absent from the dataset which means an element of survivorship bias is present in this study.

A second bias is Look Ahead Bias which occurs when information used in the study would not have been known at the time of analysis. This occurs due to the release of financial statements and the new data they provide. Financial statements are presented at a point in time, but are only released at a later date. JSE requirements state that audited financial statements must be released within 3 months of year end. This bias is mitigated through the data collection process, as features which are updated on the release of new financial information are updated at the time of release and not the financial year end date. This is essentially a timing issue within the data set, e.g. Company A has a 31 December year end. However, its results are only released to the public on the 28th of February. This means that the new information is incorporated into the share price not at the year-end but rather when the information is released at the end of February. If the dataset includes the new financial information at year end and not at release date, it will create a look-ahead bias as the historical information the algorithm is training on was not actually available at the time.

3.3.3 Limitations

The nature of the dataset is such that having a perfect dataset is unlikely. Certain values will be missing for companies in the dataset. These can take on various forms.

1. Error in the data collection process lead to a feature with no values.
2. The values exist but values for certain dates are missing.
3. The feature cannot be calculated for that company due to the nature of the business or some other “business reason.” As an example, A company does not pay a dividend and therefore will have a Dividend Yield of 0.

In order to be included in the experiment a company must have a share price with which to calculate returns from for the length of the experiment and must include all the features being tested for that experiment. Any share with missing information is excluded from that particular experiment.

Other Issues

The data is a collection of attributes with different update periods. Some attributes, e.g. share price, change weekly, while others, like ROE, only change annually on the publication of the company's Annual Financial Statements. Therefore, some features will update every period while others might only update once per year.

3.4 Experiment 1 – Factor Based Investment Strategy

3.4.1 Methodology

The first part of the experiment uses regression tests to assess the correlation between the factors and share returns. The objective of these experiments is to determine which factors present the strongest relationship to excess returns over the periods tested, and to corroborate this with the excess returns of the factors based on the portfolio sort method.

The tests were conducted as follows. Data is sorted into weekly slices of the data set. Each slice contains the forward 3-month excess return per share and the normalized factor value for the factor being tested. Excess return is calculated as the return generated by the share, less the equally weighted average return of the entire share population for that week. This can be seen as a proxy for the market performance. For each week, a regression coefficient is calculated and a time series of regression coefficients for a factor are generated. Experiments will be performed for each factor individually over the pre-crisis, crisis, post crisis and full period. For each of the 14 factors, tests are performed over each of the different time periods.

Once significant factors have been identified, the construction of the factor-based portfolios will be performed. For each week in the dataset, shares will be sorted based on the factor value. The top and bottom portfolios will be formed on the top and bottom 25% of shares. The excess return for the week is then calculated by calculating the return for each equally weighted portfolio, and the equally weighted return for the entire population of shares for that week. The difference between each portfolio and the market return is the excess return earned by that portfolio. This process is repeated for each week in a given period and the results are averaged. This provides us with the average payoff for each factor over all four time periods.

An investor implementing this strategy would do so as follows: The regression tests for each factor would be performed on the data in order to identify the most significant factors, i.e.

those with the lowest p-value. A number of these factors (in this case three) would be selected from this process. The shares on the JSE would then be sorted based on the most significant factors, the top 25% of shares would be bought to create the top portfolio and a short position would be taken in the bottom 25% of shares to create the bottom portfolio. (A short position gains when the share price drops, i.e. buy high, sell low). The payoff from these portfolios would then be realized in 12 weeks (3 months) when the portfolios would be rebalanced.

3.4.2 Evaluation Metrics

The time series of regression coefficients will be tested for significance by using a single sample students t-test, testing for a difference from a mean of 0. The p-value will be used to determine significant, with the most significant factors having the lowest p-value. These will be corroborated by the average excess returns of the same portfolios for the duration of the data. Significant factors should show excess returns, in order to demonstrate this the average excess return for a portfolio of the top and bottom 25% of shares in the investable universe will be calculated using the portfolio sort methodology. Shares will be ranked by the factor being tested, and portfolios created by placing the top 25% of shares by factor value in the top portfolio, and the bottom 25% of shares in the bottom portfolio.

3.5 Experiment 2 - Machine Learning

3.5.1 Methodology

The Machine Learning aspect of the experiment will make use of the same data and same basic methodology as the FBI experiment. The objective of this experiment is to show that SVMs are adept at portfolio selection and provide more consistent excess returns than a factor-based investment strategy based on regression and portfolio-sort. The experiment is framed as a classification task, to classify a share as an over or under performer. The probability of an observation belonging to a class is used as a proxy for the probability of a share being an over or under performer, which will be the basis of the portfolio selection rules.

A Gaussian kernel will be used as it was identified in the literature as the most suitable kernel for this method. The experiment will be treated as a classification task. The task will be to classify a share as an overperformer or an underperformer. Overperformers being shares which provide an excess return greater than 0 (Class 1) and underperformers being those which provide a return less than or equal to 0 (Class 0). The target variable will be the excess return generated by that share over the evaluation period of 3 months (12 weeks).

$$y_i = \frac{P^i_{t+3} - P^i_{t0}}{P^i_{t0}} - market\ return$$

A rolling window will be used to partition the data. The SVM will train on the first 52 weeks of data, and then predict the probability of an observation belonging to Class 1 or Class 0 for the next 22 weeks of data. It then steps forward by 22 weeks and repeats the process for the duration of the dataset. This results in every observation from week 53 to the end of the dataset having a prediction value for overperform and underperform. This gives a training/test split of 70:30.

Error!

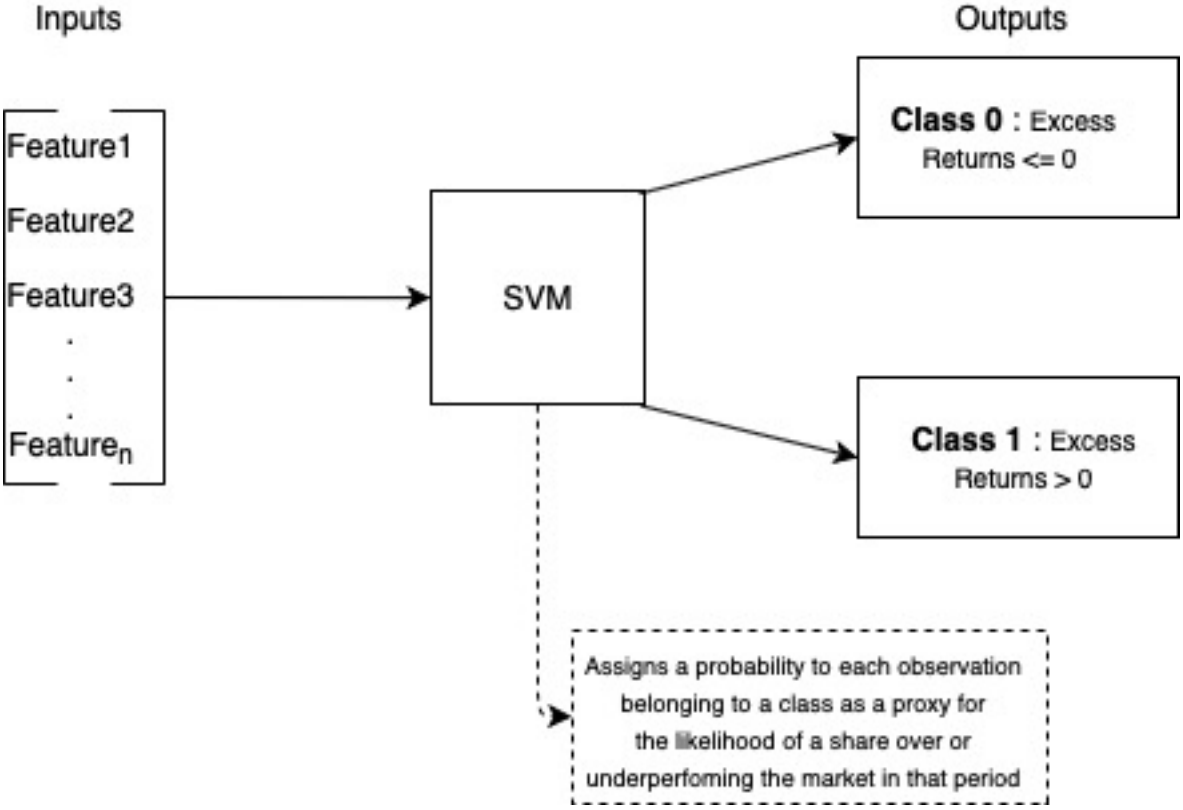

Inputs
Outputs
Feature1
Feature2
Feature3
.
.
.
Feature$_n$
SVM
Class 0 : Excess Returns <= 0
Class 1 : Excess Returns > 0
Assigns a probability to each observation belonging to a class as a proxy for the likelihood of a share over or underperfoming the market in that period

Figure 4: Diagram of SVM

The SVM will be tested over the pre-crisis, crisis, post-crisis and full periods with the following set of features (and with a range of hyperparameters):

- All features
- Quality features - $X_Q^i = \begin{bmatrix} ROE \\ ROA \\ ROIC \\ PTCF \end{bmatrix}$
- Value features - $X_V^i = \begin{bmatrix} PE \\ DY \\ PTB \\ EY \\ MVTB \end{bmatrix}$
- Liquidity and Leverage features - $X_{LL}^i = \begin{bmatrix} QR \\ DC \\ DE \\ CR \end{bmatrix}$

The above features will be used as the inputs to the SVM as depicted in Figure 4. For each of the four periods all features, quality features, value features and liquidity and leverage features will be tested. When given a set of inputs, the SVM will assign to each share a probability of it belonging to the Overperformer class (Class 1) or Underperformer class (Class 0). The shares are then ranked by probability each week and the top 25% of shares are put into the Top Portfolio and the bottom 25% into the Bottom Portfolio. The excess return for these portfolios is then calculated and thus the payoff calculated. This is done for each week, and the final result is the average excess return generated by the factor over the period being tested.

The SVM will use the rolling window methodology, however, an additional experiment will be performed using all the features and a growing window, where the size of the training set increases with each iteration.

Certain hyperparameters will need to be selected for the model. The most important of which, in an SVM, is C. C determines the level of accuracy with which the model is fitted to the training data. Too high a level of C will result in overfitting of the data, while too low a value will result in misclassification. Other parameters include gamma (γ), the width of the Gaussian kernel. This determines the sphere of influence of a training example, and can intuitively be thought of as the margin/distance between the hyperplane and a training example/observation.

3.5.2 Implementation

Data was collected for each share in individual .csv files. The 12 week forward return was calculated for each share and then the files were joined, maintaining the chronological order of the share information. The data was read into a Pandas DataFrame. The data was then normalized using the in-built packages from sci-kit learn.

The algorithm looped though the dataset, trained on x weeks of data and predicting on the next y weeks of data, then stepping forward y weeks and repeating the process. This was done for each pre-crisis, during the crisis, post-crisis and over the full period. The results from the prediction sets were gathered and averaged in order to generate the model results. This is essentially a non-random cross-validation and was done in order to preserve the temporal nature of the data. Hyper-parameter tuning was done through a grid-search, the results of each combination are presented in the tables in chapter 4 and the appendix in Tables 22, 23 and 24.

The model confidence was used as the basis for portfolio selection. The top quarter of shares with the highest model confidence of being outperformers was selected as the top portfolio and the same number of shares with the lowest model confidence where selected as the models chosen worst performers. This method was used as a proxy for portfolio-sort, whereby shares are ranked and chosen based on a factor, in order to make the methodologies comparable. The selected portfolios then have their return calculated and the payoff can then be computed in order to directly compare the two methods.

3.5.3 Evaluation Metrics

The model will be evaluated based on the model score – this is the classification accuracy of the algorithm and returns the percentage of test set observations correctly classified into the correct class [31].

It will then be compared to the Factor Based Investment Strategy on the basis of average portfolio payoff. Average portfolio payoff is the mean payoff of all the factor portfolios formed during a time period. The methodology with the higher payoff percentage can be seen as the better method.

3.6 Chapter Summary

The design of these experiments has been chosen as it will allow for the assessment of the experiments individually and allow for comparison between experiments. Experiment 1 will inform us of the best single factor portfolios through both statistical significance via the p-value and through financial performance based on payoff. These results will also allow for the construction of multi-factor portfolios. Experiment 2 will make use of feature groups, as SVMs benefit from more dimensions of data. These feature groups will be assessed through model score – which is classification accuracy – and through payoff to allow for comparison between the FBI and the ML model.

4 Results

The results for Experiment 1 and Experiment 2 are presented in the following chapter. The results for Experiment 1 consist of two tables for each experiment. The ordered regression results table, which show the factors and their p-values, and the ordered payoff results table which show the payoff made on a portfolio based on a factor. Both tables are ordered so that the most significant factor and the factor with the highest payoff are at the top. For Experiment 2, only one table is presented for each experiment. The table contains the hyperparameters used, the payoff earned per period and the model score. These tables are ordered on model score. Where comparison is made between the FBI and ML strategies, it is done through comparing the payoff. A summary of results can be found in table 16.

4.1 Pre-Crisis Period – Single Factor

Tables 4 and 5 show the results of a single factor FBI performed in the pre-crisis period. Table 4 shows the factors and the p-values obtained. The lower the p-value, the more significant a factor. Factors with low p-values should have a stronger relationship with returns. Table 5 shows the average payoff earned on a factor for the pre-crisis period. Put simply; if an investor invested R100 into a ROE based portfolio with a R50 long position and R50 short position, in 12 weeks he would have earned on average R7,41 during the pre-crisis period and his total investment would be worth R107,41.

TABLE 4: ORDERED REGRESSION RESULTS

	PRE CRISIS (2004-2007)	
	T-STAT	PVAL
PTCF	**-11.32047683**	**4.94E-22**
ROIC	7.976129043	3.11E-13
ROA	7.096560957	4.30E-11
EY	5.839044812	2.98E-08
MVTB	5.742010576	4.79E-08
MC	-5.65241816	7.40E-08
DC	-3.873655092	0.000157795
ROE	-3.691639563	0.000308139
CR	-3.371389308	0.000944299
DE	-3.249381183	0.001418425
PE	2.901180689	0.004258909
DY	-2.339240315	0.020599052
PTB	2.183155862	0.030528561
QR	-0.072425841	0.942356393

TABLE 5: ORDERED PRE-CRISIS PAYOFF

PRE CRISIS PAYOFF (%)	
ROE	**7.410865616**
ROIC	5.351594597
PTCF	5.194350249
ROA	4.42687648
CR	2.851410536
MC	2.449612666
EY	2.05962349
PTB	1.756563162
QR	1.719557457
MVTB	1.344783596
DC	1.095270443
PE	1.049540441
DE	1.048114335
DY	0.610294063

The period before the financial crisis was a bull period in the market. Confidence was high, and returns were strong, with double digit growth seen across the market. The regression experiments found PTCF, ROIC and ROA to be the three most significant factors, in terms of p-values. This would suggest that these factors would provide the best payoff over the period.

Using the information from the regression experiments, portfolios are formed based on the portfolio-sort method. The top three factors from the regression experiments yield the second to fourth best payoff. However, ROE is found to be the factor with the best payoff in this period, with a payoff of 7.41%. PTCF and EY were found to be significant factors in Muller and Ward [6] and Kruger and Toerien [5].

In Kruger's experiments on the pre-crisis stable period, he found Cashflow to Price (the inverse of PTCF), Book Value to Market (the inverse of MVTB) and size to be significant factors [5]. He also found DY to be a poor performing factor. Similar findings are present in this study where PTCF performs best and DY is among the worst performing factors identified by the regression experiments.

4.2 Crisis Period – Single Factor

The results of the FBI experiment during the financial crisis period are presented below in tables 6 and 7. Table 6 contains the regression experiment results and table 7 contains the payoff results for the period.

TABLE 6: ORDERED CRISIS PERIOD REGRESSION RESULTS

	CRISIS (2007-2009)	
	T-STAT	PVAL
ROIC	**8.52087544**	**1.31E-14**
ROE	8.4901894	1.57E-14
ROA	8.40153561	2.64E-14
PTB	-5.9025918	2.18E-08
DE	-4.1418149	5.65E-05
MVTB	3.87872347	0.00015483
QR	-3.671692	0.00033112
CR	-3.5648683	0.00048432
PTCF	-2.9487687	0.00368477
DY	2.63846597	0.00917804
EY	2.08568646	0.0386456
DC	1.11199349	0.26786271
PE	-1.0873308	0.27857856
MC	-0.2036737	0.83887562

TABLE 7: ORDERED CRISIS PERIOD PAYOFF

CRISIS PAYOFF (%)	
PTCF	**6.938193**
ROE	5.409972
MC	5.31707
ROIC	5.023663
ROA	3.926161
PTB	2.814261
MVTB	2.111925
EY	1.724631
PE	1.557328
DE	1.290544
DY	1.283607
DC	0.97817
QR	0.360445
CR	0.199549

During the financial crisis the global economy entered a prolonged recession which lead to many companies failing or experiencing declines in value. The regressions identified ROIC, ROE and ROA as the three most significant factors. These three factors all returned above average payoffs for the period and were among the top five factors during this period. During the crisis period Kruger found Cashflow to Price (the inverse of PTCF) to be the only significant factor before any adjustments [5]. While PTCF is not significant by its p-value in this study, it did produce the best payoff.

An interesting factor to look at in this period is MC (Market Capitalization) which is a proxy for the size of the company. It returned the lowest p-Value in the regression analysis but the third best payoff over the period. This could be explained by larger companies being better equipped to weather the financial crisis, whereas smaller companies may not have had the resources to survive the tough business environment. We see this in appendix table 18, as the top MC portfolio produces an average excess return of 3.354% while the bottom portfolio produces a -1.962% average excess return over the same period.

4.3 Post-Crisis Period – Single Factor

The post crisis period was the longest period tested and covers the recovery period after the financial crisis. Table 8 shows the regression results and table 9 shows the payoff results for this period.

A similar pattern to the previous periods is observed. ROIC, ROA and ROE are all significant factors identified by the regressions – as well as EY.

When looking towards the payoff, the top four factors identified in the regression experiments are among the top five factors in terms of payoff. During this period CR (Current Ratio) performed well in terms of payoff, above the performance suggested by the regression results. The appearance of CR as a high payoff factor may be due to a change in risk tolerance following the crisis. Whereby the market preferred companies that were more liquid, and less risky than in prior periods.

EY was found to be a significant factor in multiple studies [5, 6], Muller and Ward found Return on Capital to be significant [6] – which is a similar ratio to ROIC. While Kruger and Toerien identified Book to Market as significant in their study [3] – which is the inverse of MVTB.

Table 8: Ordered Post-crisis regression results

	Post Crisis (2009-2017)	
	T-Stat	pVal
ROIC	**12.1329253**	**3.33E-29**
EY	11.6681302	2.12E-27
ROA	10.1049053	1.31E-21
ROE	9.82695523	1.25E-20
MVTB	9.00076102	8.15E-18
DC	-8.3508056	1.02E-15
MC	-6.2559419	9.85E-10
PE	-5.546304	5.20E-08
DE	-5.122453	4.63E-07
PTB	-2.6795641	0.00766509
DY	-1.8433128	0.06599623
CR	-1.5659607	0.11812
QR	-0.0876196	0.93022127
PTCF	-0.0546006	0.95648291

Table 9: Ordered Post crisis payoff

Post Crisis Payoff (%)	
ROE	**9.41175931**
EY	5.84387838
ROIC	5.67952612
CR	5.50327208
ROA	4.92759176
QR	3.79956036
PTCF	3.4505074
DE	3.3867726
DC	3.1687109
PTB	0.69886552
MVTB	0.57764238
PE	0.54175243
MC	0.41784468
DY	0.06301936

4.4 Full Period Results – Single Factor

The final FBI experiment was performed over the entire period of the data. This experiment does not make distinction of market conditions but gives an idea of the stylized facts that persist most strongly throughout the market. Table 10 contains the regression results and Table 11 the average payoff for a 12-week holding period.

TABLE 10: ORDERED FULL PERIOD REGRESSION RESULTS

	FULL PERIOD (2004-2017)	
	T-STAT	PVAL
ROIC	**16.5369068**	**2.21E-52**
ROA	14.4923694	5.36E-42
EY	12.5275592	9.70E-33
MVTB	11.1729926	7.38E-27
ROE	11.0799338	1.80E-26
DE	-7.0195703	5.11E-12
PTCF	-6.6946936	4.32E-11
MC	-5.9702836	3.70E-09
DC	-5.7922475	1.03E-08
CR	-3.577728	0.00036958
PTB	-2.916581	0.00364792
QR	-1.5920592	0.11180609
DY	-0.9754024	0.32968514
PE	0.30253709	0.7623292

TABLE 11: ORDERED FULL PERIOD PAYOFF

FULL PERIOD PAYOFF (%)	
ROIC	**5.44850997**
ROA	5.42285738
PTCF	5.204043
ROE	5.15052395
EY	4.4690427
MC	3.17952786
PTB	2.0154866
PE	1.90276825
MVTB	1.43219609
DE	1.12124767
CR	1.02825622
QR	0.89365565
DY	0.80612657
DC	0.77538519

The full period regression results give an idea of the consistency of returns of the factors, irrespective of market conditions. The three most significant factors identified by the regression experiment are ROIC, ROA and EY. ROIC and ROA are the two best performing factors in terms of payoff over the full period, while EY is fifth. The two factors that complete the top five in terms of payoff are PTCF and ROE. Both of these factors have consistently performed well when measured in terms of payoff over all the periods tested.

ROIC and ROA were not included in Kruger's study, however he found Cashflow to Price (inverse of PTCF), Book to Market Value (the inverse of MVTB), Size and EY to be significant factors over the full period when using unadjusted data [5].

From the evidence presented, there is a link between the regression results, and the resulting payoff of the factor when used to make an investment decision.

4.5 *SVM Results and Comparison*

The results for the SVM based investment strategy and the relevant comparisons to an FBI are presented below. Table 12 contains the results for a multi-factor FBI and table 13 contain the results for the multi-factor SVM experiments, tables 16 and 17 contain the results for the all-factor SVM experiments and table 17 presents a summary of the comparison of results for the different strategies.

A multi-factor portfolio relies on a combination of related factors in order to create a portfolio. These were chosen to be Quality, which informs us about the quality of a company's earnings and financial position, Liquidity and leverage which is looks at a company's long and short-term debts, and its ability to finance these debts and Value which is a collection of ratios that describe a shares value relative to its price.

For the Factor Based Investment strategy, the portfolios were formed by combining and averaging the individual constituents of that portfolio. E.g.: The Quality portfolio is formed by taking the top portfolios for ROE, ROIC, ROA and PTCF and averaging the return of that portfolio, repeating the process for the bottom portfolio and using the absolute difference of the two values to calculate the payoff for a particular period.

TABLE 12: FBI MULTI-FACTOR PORTFOLIO PAYOFF

FACTOR BASED INVESTMENT STRATEGY	PRE - CRISIS PAYOFF (%)	CRISIS PAYOFF (%)	POST CRISIS PAYOFF (%)	FULL PERIOD (%)
QUALITY *ROE, ROIC, ROA, PTCF*	2.99874661	1.85540077	4.14209245	5.30648357
LIQUIDITY AND LEVERAGE *QR, DE, DC, CR*	0.606895804	0.52695437	0.686837237	0.95463618
VALUE *PE, DY, PTB, EY, MVTB*	1.36416095	1.20849803	1.51982387	2.12512404

TABLE 13: SVM MULTI-FACTOR RESULTS

SVM C=10, γ=0.5	PRE - CRISIS PAYOFF (%)	CRISIS PAYOFF (%)	POST CRISIS PAYOFF (%)	FULL PERIOD PAYOFF (%)	CLASSIFICATION ACCURACY
QUALITY *ROE, ROIC, ROA, PTCF*	3.4238273	3.3971757	2.8136632	3.0428695	0.56533249
LIQUIDITY AND LEVERAGE *QR, DE, DC, CR*	0.7390215	5.8714583	1.52631014	2.4105315	0.56769579
VALUE *PE, DY, PTB, EY, MVTB*	1.673575	2.554708	3.85576304	3.21791115	0.56922121

The SVM results selected are for those with the highest average classification accuracy of the three different portfolios for a given set of hyperparameters and not necessarily of the hyper parameters which gave each individual portfolio the highest score. The portfolios were formed by providing the constituents of each category as the input features into the SVM. The SVM calculated the probability of an observation belonging to the over performer class, and the ranked list was used to construct the portfolios. The top portfolio containing the 25% of shares considered most likely to be over performers and the bottom portfolio consisting of the 25% of shares predicted to have the lowest possibility of being an overperformer.

When analysing the results per period it becomes clear that the SVM outperforms the FBI. The SVM selected portfolios produce better payoffs in ten out of the twelve tests (3 by category and 4 by time period), only being beaten by the Quality portfolio in the Post Crisis and Full Period tests. However, in other tests, the SVM significantly outperforms the FBI portfolios. In particular, the liquidity and leverage portfolio during the crisis period where the SVM produced a 5.34% greater payoff than the FBI portfolios.

Runhaar did not account for market conditions in his study, and conducted his study over the full period. He found that DY, EY, and ROE all produced greater returns when used by an SVM compared to an FBI [2]. Runhaar found the momentum multi-factor portfolio to provide the lowest returns of the multi-factor portfolios tested [2]. He found quality to be the next best multi-factor portfolio, followed by value. This is in line with these findings where value is the best performing multi-factor portfolio, followed by quality. Runhaar also tested a multifactor portfolio consisting of all the factors and found this to be the best performing of all [2]. This is also consistent with the findings presented in Table 14, which shows all factor portfolios outperform the multi-factor portfolios over all periods.

4.6 All Factor SVM Results

TABLE 14: ALL FACTOR SVM RESULTS

HYPER PARAMETERS	PRE-CRISIS PAYOFF (%)	CRISIS PAYOFF (%)	POST CRISIS PAYOFF (%)	FULL PERIOD PAYOFF (%)	CLASSIFICATION ACCURACY
C = 200, γ = 0.05	**6.6821709**	**6.1292763**	**8.1420912**	**7.4509475**	**0.61781072**
C = 500, γ = 0.05	6.0707878	5.24946684	7.8004295	6.9431102	0.61471197
C = 100, γ = 0.1	6.3961253	5.77317055	7.4972617	6.9283056	0.61308173
C = 50 γ= 0.1	6.6367727	5.8992467	7.772112	7.1634438	0.61294506
C = 10, γ = 0.1	7.3446104	6.399997	7.4876253	7.2138217	0.60782828
C = 200, γ = 0.1	6.241176	5.5866406	7.0506921	6.5869212	0.6075925
C = 10, γ = 0.5	5.7827242	4.44765236	6.1661989	5.7092628	0.59344178
C = 50 γ = 0.5	5.1053843	3.95273372	5.2061172	4.9004729	0.58990512
C = 1, γ = 0.1	6.0116329	6.11794968	7.6880069	7.0659413	0.58839929

When using all factors as input features into the SVM there is a notable jump in payoff compared to both the SVM using categorical features and the FBI strategy. The all factor SVM produced a payoff of 6.68% in the pre-crisis period, 6,12% during the financial crisis, 8,14% in the post crisis period and a full period payoff of 7.45%. The payoff generated for each period was greater than the payoff for any other multi-factor strategy tested. It also generated a better payoff than all single factor strategies other than ROE in the pre-crisis period which returned 7.41% payoff, PTCF in the crisis period which returned 6.93% payoff and ROE in the post crisis period, which provided a payoff of 9.41%. Using all the factors also improved the model score compared to the multi-factor SVM.

4.7 All Factor Growing Window SVM Results

TABLE 15: SVM GROWING WINDOW RESULTS

HYPER PARAMETERS	PRE-CRISIS PAYOFF (%)	CRISIS PAYOFF (%)	POST CRISIS PAYOFF (%)	FULL PERIOD PAYOFF (%)	CLASSIFICATION Accuracy (%)
C = 10 γ = 0.05	**7.2484317**	**11.551833**	**12.5365513**	**11.4926626**	**0.65660702**
C = 10, γ = 0.1	**8.1506429**	**9.883608**	**12.3550892**	**11.1342986**	**0.65567775**
C = 25, γ = 0.1	8.3388347	9.4109322	11.9010345	10.775033	0.64815343
C = 100, γ = 0.1	7.5287785	7.6544515	10.792774	9.562754	0.63869738
C = 1, γ = 0.1	7.0255974	10.8867321	11.1816659	10.4721096	0.63801677
C = 10, γ = 0.25	7.2067456	7.0100623	10.1254864	8.9540393	0.62943833
C = 1, γ = 0.5	5.8082905	7.56256943	8.8063476	8.055862	0.6190433

The final experiment performed involved using all the factors to train the SVM, but instead of rolling the window through the periods, the initial start of the window remained constant and only the size of the training set increased with each loop. This resulted in the SVM being able to train for longer and this produced the best results of all the experiments – as shown in table 15. The hyperparameters that produced the best score were C = 10, γ = 0.05. However, other parameters may have provided better payoff over certain periods.

Using the results from the experiment with the highest model score, this method was only outperformed by the single factor ROE portfolio in the pre-crisis period (7.41%). The difference between payoffs is negligible however, at only 0.16%. Over all other periods, this experiment provided the best payoff and highest model scores.

Table 16: Full Results comparison Table

	Pre-Crisis (%)	Crisis (%)	Post-Crisis (%)	Full Period (%)
SVM: Quality	3.4238273	3.3971757	2.8136632	3.0428695
FBI: Quality	2.99874661	1.85540077	4.14209245	5.306483574
SVM - FBI	**0.42508069**	**1.54177493**	**-1.3284293**	**-2.263614074**
SVM: Liquidity and Leverage	0.7390215	5.8714583	1.52631014	2.4105315
FBI: Liquidity and Leverage	0.606895804	0.526954371	0.68683724	0.954636183
SVM - FBI	**0.132125696**	**5.344503929**	**0.8394729**	**1.455895317**
SVM: Value	1.673575	2.554708	3.85576304	3.21791115
FBI: Value	1.36416095	1.20849803	1.51982387	2.125124043
SVM - FBI	**0.30941405**	**1.34620997**	**2.33593917**	**1.092787107**
SVM: All Factors	6.6821709	6.1292763	8.1420912	7.4509475
FBI: Avg All	2.740604081	2.781108617	3.39076452	2.774973365
SVM - FBI	**3.941566819**	**3.348167683**	**4.75132668**	**4.675974135**
SVM: Growing window	7.2484317	11.551833	12.5365513	11.4926626
FBI: Avg All	2.740604081	2.781108617	3.39076452	2.774973365
SVM - FBI	**4.507827619**	**8.770724383**	**9.14578678**	**8.717689235**

The results of this study show a clear benefit to using an SVM over a traditional factor-based investment strategy. Overall an SVM based strategy provided a greater and more consistent payoff than FBI. The SVM performed better in all but two of the tests. The best results achieved by the SVM were with all the features included, and with the longest training period. This suggests that better results could be achieved with the use of more factors and with the use of the growing window method.

Conclusion

This study aimed to compare the results of a factor-based investment strategy implemented using traditional techniques, against a machine learning implementation. It focused on the formation of portfolios based on companies underlying financial ratios as factors and used the payoff, i.e. the positive return on the top portfolio minus the negative return on the bottom portfolio, as the metric to evaluate and compare the two strategies.

It found that a factor-based investment strategy can produce excess returns over the market, in line with previous studies of this nature. However, when implementing a Support Vector Machine in a classification task, greater and more consistent excess returns could be attained. It found that the Machine Learning method benefitted from increasing the number of features used, for a given training window size. It also found that increasing the size of the training window further improved results.

Experiments were performed over different market periods, which identified that certain factors become more significant during different market periods. This information is valuable in allowing for the adjustment of portfolio composition during different market periods. Factors such as PTCF and MC were shown to be more important when the market is in a contractionary period, while ROE, ROIC and ROA are shown to be significant factors during expansionary market periods.

This study ignored certain realities in building the model, the most significant of which are the costs associated with share trading and portfolio rebalancing. These material costs can influence the margins earned on investments as they are incurred every time shares are traded, which has the effect of reducing profits. This effect is more significant when dealing with smaller investments.

Future studies may include different algorithms other than an SVM. The holding period can be adjusted and the inclusion of a larger number of shares and a larger number of factors may also contribute to improved performance.

Bibliography

[1] K. Kyoung-jae, "Financial time series forecasting using Support Vector Machines," Dongguk University, 2003.

[2] A. J. Runhaar, "Active portfolio management: Improving factor-based portfolio construction by applying machine learning to classify stock performance," UCT Graduate School of Business, 2016.

[3] B. G. Malkiel, "The Efficient Market Hypothesis and Its Critics," *The Journal of Economic Perspectives,* vol. 17, no. 1, pp. 59-82, 2003.

[4] M. V. Sewell, "History of the Efficient Market Hypothesis," University College of London2011.

[5] R. Kruger and F. Toerien, "The Consistency of Equity Style Anomalies on the JSE During a Period of Market Crisis," *The African Finance Journal,* vol. 16, no. 1, 2014.

[6] C. Muller and M. Ward, "Style-based effects on the Johannesburg Stock Exchange: A graphical time-series approach," *Investment Analysts Journal,* no. 77, 2013.

[7] P. van Rensburg, "A decomposition of style-based risk on the JSE," *Investment Analysts Journal,* vol. 54, 2001.

[8] P. van Rensburg and M. Robertson, "Style characteristics and the cross-section of JSE returns," *Investments Analysts Journal,* vol. 32, no. 57, pp. 7-15, 2003.

[9] M. V. Sewell, "Application of Machine Learning to Financial Time Series Analysis," Doctor of Philosophy, Department of Computer Science, University College of London, 2017.

[10] E. F. Fama and J. D. MacBeth, "Risk, return and equilibrium: Empirical tests," *Journal of politcal economy,* vol. 81, no. 3, pp. 607-636, 1973.

[11] C. Auret and R. Sinclaire, "Book-to-market ratio and returns on the JSE," *Investment Analysts Journal,* vol. 35, no. 63, pp. 31-38, 2006.

[12] A. Hoffman, "Stock return anomalies: Evidence from the Johannesburg Stock Exchange," *Investment Analysts Journal,* no. 75, 2012.

[13] E. F. Fama and K. R. French, "Dissecting Anomalies," *The Journal of Finance,* vol. 63, no. 4, pp. 1653-78, 2008.

[14] L. Underhill and D. Bradfield, *Introstat*. Department of Statistical Sciences: University of Cape Town, 2013.

[15] R. Kruger, "Asset Pricing: Style-Based Risk on the JSE," ed: UCT, 2016.

[16] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development,* vol. 3, no. 3, 1959.

[17] K. P. Shung. (2018, 30 November). *Accuracy, Precision, Recall or F1?* Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[18] R. L. Wilson and R. Sharda, "Bankruptcy prediction using neural networks," *Decision support systems,* vol. 11, no. 5, pp. 545-557, 1994.

[19] J. H. Min and Y.-C. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters.," *Expert systems with applications,* vol. 28, no. 4, pp. 603-614, 2005.

[20] Y.-C. Lee, "Application of support vector machines to corporate credit rating prediction," *Expert systems with applications,* vol. 33, no. 1, pp. 67-74, 2007.

[21] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study.," *Decision support systems,* vol. 37, no. 4, pp. 543-558, 2004.

[22] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science,* vol. 2, no. 1, 2011.

[23] L. Cao and F. E. H. Tay, "Financial Forecasting Using Support Vector Machines," 2001.

[24] X.-Y. Qian and S. Gao, "Financial Series Prediction: Comparison Between Precision of Time Series Models and Machine Learning Methods," 2017.

[25] N. Milosevic, "Equity forecast: Predicting long term stock price movement using machine learning," 2016.

[26] A. Fan and M. Palaniswami, "Stock Selection using Support Vector Machines," Department of EEE, University of Melbourne, 2001.

[27] S. Arik, B. Eryilmaz, and A. Goldberg, "Supervised classification-based stock prediction and portfolio optimization," 2014.

[28] V. V. C Cortes, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995 1995.

[29] R. Gandhi. (2018, 09 October 2019). *Support Vector Machine - Introduction to Machine Learning Algorithms*. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[30] C. Albon. (2017, 09 October). *SVC Parameters WHEN Using RBF Kernel*. Available: https://chrisalbon.com/machine_learning/support_vector_machines/svc_parameters_using_rbf_kernel/

[31] Scikit-Learn. (2017). Available: http://scikit-learn.org/stable/

6 Appendix

TABLE 17: PRE-CRISIS PORTFOLIO PERFORMANCE

Pre Crisis			
	Top Portfolio %	Bottom Portfolio %	Payoff %
PE	0.548198789	-0.501341651	1.04954044
DY	0.794171954	0.183877891	0.61029406
ROE	-3.0731865	4.337679117	7.41086562
QR	0.916766046	-0.802791411	1.71955746
ROIC	-2.59481489	2.756779702	5.3515946
DE	-0.1545742	0.89354013	1.04811433
PTB	1.457130738	-0.299432423	1.75656316
PTCF	2.733357651	-2.460992599	5.19435025
ROA	-1.85122514	2.575651341	4.42687648
EY	1.237215578	-0.822407912	2.05962349
MVTB	1.104322593	-0.240461003	1.3447836
MC	1.433375203	-1.016237463	2.44961267
DC	-0.08767912	1.007591318	1.09527044
CR	2.297648907	-0.553761629	2.85141054

TABLE 18: CRISIS PERIOD PORTFOLIO PERFORMANCE

Crisis			
	Top Portfolio %	Bottom Portfolio %	Payoff %
PE	0.865029059	-0.692299395	1.55732845
DY	1.195628176	-0.08797931	1.28360749
ROE	-2.27974693	3.130224997	5.40997192
QR	-0.26572314	0.094722308	0.36044545
ROIC	-1.87707973	3.146583346	5.02366307
DE	0.662000863	-0.628543069	1.29054393
PTB	2.375207255	-0.439053547	2.8142608

PTCF	3.709362893	-3.228830207	6.9381931
ROA	-0.81992359	3.106237603	3.9261612
EY	-0.25434385	1.470287552	1.7246314
MVTB	1.844058702	-0.26786611	2.11192481
MC	3.354781705	-1.962288307	5.31707001
DC	0.699151627	-0.279018385	0.97817001
CR	-0.4375823	-0.637131293	0.19954899

TABLE 19: POST CRISIS PORTFOLIO PERFORMANCE

Post Crisis			
	Top Portfolio %	Bottom Portfolio %	Payoff %
PE	0.23136852	-0.310383908	0.54175243
DY	0.392715733	0.455735093	0.06301936
ROE	-3.86662607	5.545133238	9.41175931
QR	2.099255233	-1.70030513	3.79956036
ROIC	-3.31255006	2.366976059	5.67952612
DE	-0.97114927	2.415623329	3.3867726
PTB	0.539054221	-0.1598113	0.69886552
PTCF	1.757352408	-1.693154991	3.4505074
ROA	-2.88252668	2.045065079	4.92759176
EY	2.728775009	-3.115103375	5.84387838
MVTB	0.364586483	-0.213055896	0.57764238
MC	-0.4880313	-0.070186618	0.41784468
DC	-0.87450988	2.29420102	3.1687109
CR	5.032880118	-0.470391966	5.50327208

Table 20: Full Period Portfolio Performance

Full Period			
	Top Portfolio %	Bottom Portfolio %	Payoff %
PE	1.17147445	-0.731293798	1.90276825
DY	0.60936424	-0.196762328	0.80612657
ROE	-3.14182372	2.008700234	5.15052395
QR	-1.34953914	-0.455883494	0.89365565
ROIC	-3.37428424	2.074225722	5.44850997
DE	-0.19618314	-1.317430802	1.12124767
PTB	1.86413359	-0.151353015	2.0154866
PTCF	3.19142569	-2.01261731	5.204043
ROA	-3.4973522	1.925505182	5.42285738
EY	-2.45809818	2.010944519	4.4690427
MVTB	1.50066439	0.068468298	1.43219609
MC	2.07385824	-1.105669629	3.17952786
DC	-0.11151042	-0.88689561	0.77538519
CR	0.07065463	-0.957601589	1.02825622

Table 21: Portfolio Performance for multi-factor regression strategy

Quality			
	ROE, ROIC, ROA, PTCF		
	Top Portfolio %	Bottom Portfolio %	**Payoff %**
Pre Crisis	-1.19646722	1.802279391	**2.99874661**
Crisis	-0.31684684	1.538553935	**1.85540077**
Post Crisis	-2.0760876	2.066004846	**4.14209245**

Liquidity and Leverage			
	QR, DE, DC, CR		
	Top Portfolio %	Bottom Portfolio %	**Payoff %**
Pre Crisis	0.74304041	0.1361446	**0.6068958**
Crisis	0.16446176	-0.3624926	**0.52695437**
Post Crisis	1.32161905	0.63478181	**0.68683724**

Value			
	PE, DY, PTB, EY, MVTB		
	Top Portfolio %	Bottom Portfolio %	**Payoff %**
Pre Crisis	1.02820793	-0.335953	**1.36416095**
Crisis	1.20511587	-0.0033822	**1.20849803**
Post Crisis	0.85129999	-0.6685239	**1.51982387**

Table 22: SVM Multi Factor Strategy portfolio performance

C = 10, g = 0.5	**Quality**		
	ROE, ROIC, ROA, PTCF		
	Top Portfolio %	Bottom Portfolio %	Payoff %
Pre-Crisis	7.085883	3.6620557	3.4238273
Crisis	-0.9337068	-4.3308825	3.3971757
Post Crisis	1.0085682	-1.805095	2.8136632
Full Period	1.4967667	-1.5461028	3.0428695
Score			0.56533249

	Liquidity and Leverage		
	QR, DE, DC, CR		
C = 10, g = 0.5	Top Portfolio %	Bottom Portfolio %	Payoff %
Pre-Crisis	3.9427342	3.2037127	0.7390215
Crisis	1.2016553	-4.669803	5.8714583
Post Crisis	-0.0988018	-1.6251119	1.52631014
Full Period	0.825814	-1.5847175	2.4105315

Score		0.567695793

	Value		
	PE, DY, PTB, EY, MVTB		
C = 10, g = 0.5	Top Portfolio %	Bottom Portfolio %	Payoff %
Pre-Crisis	6.00202	4.328445	1.673575
Crisis	-1.8786993	-4.4334073	2.554708
Post Crisis	0.58374584	-3.2720172	3.85576304
Full Period	0.84985715	-2.368054	3.21791115
Score			0.56922121

Table 23: SVM All Factor Portfolio Performance (Rolling Window)

	Pre-Crisis		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	8.579143	2.5675101	**6.0116329**
C = 10, g = 0.1	8.660981	1.3163706	**7.3446104**
C = 50 g = 0.1	8.187083	1.5503103	**6.6367727**
C = 50 g = 0.5	7.516653	2.4112687	**5.1053843**
C = 10, g = 0.5	7.750701	1.9679768	**5.7827242**
C = 100, g = 0.1	8.1595745	1.7634492	**6.3961253**
C = 200, g = 0.1	8.046418	1.805242	**6.241176**
C = 200, g = 0.05	8.287895	1.6057241	**6.6821709**
C = 500, g=0.05	7.9021444	1.8313566	**6.0707878**

	Crisis		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	0.15093918	-5.9670105	**6.11794968**
C = 10, g = 0.1	0.4426205	-5.9573765	**6.399997**
C = 50 g = 0.1	0.6473387	-5.251908	**5.8992467**
C = 50 g = 0.5	-0.3166888	-4.2694225	**3.95273372**
C = 10, g = 0.5	-0.1851611	-4.6328135	**4.44765236**
C = 100, g = 0.1	0.64467585	-5.1284947	**5.77317055**
C = 200, g = 0.1	0.6495506	-4.93709	**5.5866406**
C = 200, g = 0.05	0.821703	-5.3075733	**6.1292763**
C = 500, g=0.05	0.57229984	-4.677167	**5.24946684**

	Post crisis		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	3.0512319	-4.636775	**7.6880069**
C = 10, g = 0.1	2.8142626	-4.6733627	**7.4876253**
C = 50 g = 0.1	3.150439	-4.621673	**7.772112**
C = 50 g = 0.5	1.8172095	-3.3889077	**5.2061172**
C = 10, g = 0.5	2.186566	-3.9796329	**6.1661989**
C = 100, g = 0.1	2.9720547	-4.525207	**7.4972617**
C = 200, g = 0.1	2.6901307	-4.3605614	**7.0506921**
C = 200, g = 0.05	3.2120092	-4.930082	**8.1420912**
C = 500, g=0.05	3.119013	-4.6814165	**7.8004295**

	Full Period		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	3.2329187	-3.8330226	**7.0659413**
C = 10, g = 0.1	3.1675007	-4.046321	**7.2138217**
C = 50 g = 0.1	3.348254	-3.8151898	**7.1634438**
C = 50 g = 0.5	2.2027493	-2.6977236	**4.9004729**

C = 10, g = 0.5	2.496181	-3.2130818	**5.7092628**
C = 100, g = 0.1	3.2338216	-3.694484	**6.9283056**
C = 200, g = 0.1	3.0443199	-3.5426013	**6.5869212**
C = 200, g = 0.05	3.4419858	-4.0089617	**7.4509475**
C = 500, g=0.05	3.2676158	-3.6754944	**6.9431102**

TABLE 24: SVM ALL FACTOR PORTFOLIO PERFORMANCE (GROWING WINDOW)

	Pre-Crisis		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	9.066752	2.0411546	7.0255974
C = 10, g = 0.1	9.169932	1.0192891	8.1506429
C = 100, g = 0.1	8.810378	1.2815995	7.5287785
C = 1, g = 0.5	7.9297028	2.1214123	5.8082905
C = 25, g = 0.1	9.0964575	0.7576228	8.3388347
C = 10, g = 0.25	8.4762745	1.2695289	7.2067456

	Crisis		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	3.3470607	-7.5396714	10.8867321
C = 10, g = 0.1	1.8839365	-7.9996715	9.883608
C = 100, g = 0.1	1.3328625	-6.321589	7.6544515
C = 1, g = 0.5	0.69980943	-6.86276	7.56256943
C = 25, g = 0.1	1.5231549	-7.8877773	9.4109322
C = 10, g = 0.25	0.6220263	-6.388036	7.0100623

	Post crisis		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	4.7536674	-6.4279985	11.1816659
C = 10, g = 0.1	5.1506705	-7.2044187	12.3550892
C = 100, g = 0.1	4.193284	-6.59949	10.792774
C = 1, g = 0.5	3.5605361	-5.2458115	8.8063476
C = 25, g = 0.1	4.903473	-6.9975615	11.9010345

C = 10, g = 0.25	3.9263554	-6.199131	10.1254864

	Full Period		
	Top Portfolio %	Bottom Portfolio %	Payoff %
C = 1, g = 0.1	5.093623	-5.3784866	10.4721096
C = 10, g = 0.1	5.014755	-6.1195436	11.1342986
C = 100, g = 0.1	4.243658	-5.319096	9.562754
C = 1, g = 0.5	3.5725837	-4.4832783	8.055862
C = 25, g = 0.1	4.768073	-6.00696	10.775033
C = 10, g = 0.25	3.86362	-5.0904193	8.9540393

TABLE 25: SUMMARY OF LITERATURE

Short Term Index Movement Prediction					
Paper	Algorithms Used	Performance	Target Variable and Features	Evaluation Measures	Data Set
Cao and Tay (2001)	SVM (Gaussian Kernel) MLP with BP	Finds SVMs perform better than MLP due to the Structural Risk Minimization Principle. SVMs were also faster to train.	Target Variable = Direction of price movement Features = • Four lagged RDP values based on 5-day periods, and the transformed closing price which is obtained by subtracting the 15-day moving average to eliminate trend Three technical indicators: • The Moving Average Convergence Divergence (MACD) • On Balance Volume • Volatility	Normalized Mean Squared Error Mean Absolute Error Directional Symmetry Correct Up-trend Correct Down-trend	S&P 500 Daily Data from the Chicago Mercantile Exchange 01 April 1993 – 31 December 1995
Kim (2003)	SVM (Polynomial Kernel and	SVM = 57.8313% BP = 54.7332%	Target Variable = Direction of daily price change of a stock index	Hit Ratio	KOSPI from January 1989 to

	Gaussian radial basis function) 3-layer NN with BP CBR	CBR = 51.9793%	Features = • %K – Stochastic %K compares where a security's price closed relative to its price range over a given time period • %D – Moving average of %K • Slow %D – Moving average of %D • Momentum – Price change over a given time span • ROC – Price rate of change, the difference between the current price and the price *n* days ago • Williams %R – Momentum indicator that measures overbought/oversold levels • A/D oscillator – Momentum indicator that associates change in price • Disparity5 – Distance between current price and the moving average of 5 days • Disparity10 – 10-day disparity		December 1998 (2928 trading days)

			• OSCP – price oscillator, the difference between two moving averages of a securities price • CCI – Commodity Channel Index, measures the variation of a securities price from its statistical mean • RSI – Relative Strength Index		
Qian and Gao (2017)	DAE-SVM SVM (Kernel not specified) Logistic Regression ARIMA MLP	DAE – SVM = 69.1% SVM = 64.2% LR = 62.3% ARIMA = 59.3% MLP = 56.6%	Target Variable = Closing price movement direction of the next trading day Features = • Open Price • Close Price • High Price • Low Price • Trading Volume Technical Indicators • %K • %D • Slow %D • Momentum	Hit Ratio	Weekly data from S&P 500, Dow 30 and Nasdaq indices between January 2012 and December 2016

			• ROC • Williams %R • A/D Oscillator • Disparity5 • Disparity10 • OSCP • CCI • RSI		

Long Term Market Outperformance

Paper	Algorithm Used	Performance	Target Variable and Features	Evaluation Measures	Data Set
Fan and Palaniswami (2001)	SVM (Gaussian Kernel) Portfolio Sort	SVM outperformed the market producing a 208% return over 5 years against 71% by the market benchmark.	Target Variable = Classify stock as ‘exceptional high return stock (Class +1)’ and ‘Unexceptional return (Class -1)’ Features = Return on Capital • Profit before Tax / Total Assets • Profit before Tax / Total Capital • Net Income / Total Capital • Cash flow / Total Assets • Cash flow / Total Capital Profitability • Profit before tax / Sales • Profit after tax / Sales • Net Income / Sales • Cash flow / Sales	Compares portfolio results to benchmark performance.	Stocks on the Australian Stock Exchange from 1992-2000 *Sliding Window*

			• Profit after Tax / Equity • Cash flow / Total Market Value • Profit after Tax / Cash flow Leverage • Debt / Equity • Total Liabilities / Total Capital • Total Liabilities / Shareholders Equity • Total Assets / Shareholders Equity • Total Assets / Total Market Value Return on Investment • Return on Assets Investment • PE Ratio • Net tangible assets per share • Dividend Yield • Earnings Yield • Shareholders Equity / Total Market Value		

			Growth • Sales Growth • Earnings before tax growth • Earnings after tax growth • Net recurring profit growth • Operating profit growth • Shareholders fund growth • Total Assets Growth Short Term Liquidity • Current Assets / Current Liabilities • Current Liabilities / Total Assets • Current Liabilities / Equity • Long Term Debt / Total Debt Risk • Profit before Tax / Current Liabilities • Profit After Tax / Current Liabilities • Cash flow / Current Liabilities		

Arik, Eryilmaz and Goldberg (2014)	SVM (Gaussian Kernel)	Training: Average true classification = 71.2% for bullish stocks and 60.2% for bearish stocks. Prediction: 58.8% for bullish stocks and 57.9% for bearish stocks.	Target Variable = Classify stock as “bullish” (outperforms market) or “bearish” (underperforms market) Features = 69 Fundamental financial parameters, reduced to 52 after data preprocessing. The specific parameters are not provided in the paper.	Classification Accuracy	1012 stocks listed on the NYSE over a 10-year period (2004-2013)
Milosevic (2016)	Random Forest C4.5 decision trees SVM with SMO JRip Random Tree Logistic Regression Naïve Bayes	With all features, the most accurate results were achieved by the Random Forrest Algorithm (F-Score = 75.1%). When the number of features was reduced, they found that Book	Target Variable = Predict if stock price will be 10% higher over a year period. Features = • Book Value • Market Capitalization • Change of stock Net price over the one-month period • Percentage change of Net price over the one-month period • Dividend Yield • Earnings Per Share	Precision Recall and F-Score. Tested if the algorithm could correctly classify whether a stocks value would have a 10% higher value over a 1-year period.	Quarterly stock prices of 1739 stocks from the S&P 1000, FTSE 100 and S&P Europe 350 from 2012 until 2015.

	Bayesian Networks	Value, Market Cap, Dividend yield, EPS, PE ratio, Price to book ratio, DPS, Current Ratio, Quick Ratio, Total Debt to Total Equity and historic price were the most significant features and increased the Random Forrest F-Score to 76.5%.	• Earnings per Share growth • Sales Revenue Turnover • Net Revenue • Net Revenue growth • Sales growth • PE Ratio • PE Ratio five-year average • Price to book Ratio • Price to Sales Ratio • Dividend per Share • Current Ratio • Quick Ratio • Total Debt to Equity • Analyst Ratio • Revenue growth adjusted by 5-year annual growth ratio • Profit Margin • Operating Margin • Asset Turnover		

Runhaar (2016)	SVM (Linear Kernel) Portfolio - sort	Finds SVMs out performed a traditional Factor Based Investment Strategy for single factor as well as multi-factor models.	Target Variable = percentage change in closing share price, one month in the future. Features = • Open Price • Close Price • Volume • High Price • Low Price • Dividend Yield • Earnings Yield • Market Capital * • Book Value • Return on Equity • Return on Invested Capital Portfolios constructed based on: • Value: Dividend Yield, Earnings Yield, Book to Market ratio	Compares returns to an equally weighted portfolio of shares. Uses the Sharpe Ratio as a measure of risk adjusted returns and a t-statistic for significance.	100 stocks listed on the JSE main board, using 15 years (March 2001 – February 2016) of data.

			• Momentum: 6 and 12-month Momentum • Quality: ROE and ROIC *in paper as market capital but refers to market capitalization		