TÉCNICO LISBOA

Portfolio Optimization using Fundamental Analysis with a Logistic Regression Approach

Francisco Marques Cruz Rodrigues

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisor: Prof. Rui Fuentecilla Maia Ferreira Neves

Examination Committee

Chairperson: Prof. Teresa Maria Sá Ferreira Vazão Vasques
Supervisor: Prof. Rui Fuentecilla Maia Ferreira Neves
Member of the Committee: Prof. João Paulo Baptista de Carvalho

November 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

To professor Rui Neves, for all the patience, feedback and understanding of the subject. Without it, all the problems and decisions faced would have been much harder.

To my parents and brother, who have always been there for me and supported me unconditionally.

To all my friends and colleagues, who have been part of this journey and with whom i had the pleasure of spending these last 5 years of university.

Resumo

Este trabalho propõe uma abordagem que combina modelos de Regressão Logística com Análise Fundamental, a fim de criar um sistema capaz de prever o crescimento a longo prazo das empresas, ao fazer uso das classificações percentuais do algoritmo. A implementação proposta retira informações financeiras das empresas integrantes do índice S&P 500, de 2009 a 2021. Essas informações são filtradas e transformadas em rácios financeiros, capazes de quantificar o desempenho das empresas, que serão depois utilizados pelos modelos. De seguida, os modelos calculam pontuações de probabilidade que refletem o nível de confiança de que cada empresa crescerá mais do que o índice de mercado, para cada ano de teste. As empresas são então ordenadas de acordo com seus respectivos valores de probabilidade. Os retornos anuais foram avaliados calculando o Retorno ao Investimento, Max Drawdown e Sharpe ratio, para analisar os lucros gerados, a volatilidade e o risco associado aos portfolios escolhidos. Ao longo dos 8 anos de investimento, os modelos de Máquinas de Vetores de Suporte (SVM), criados como ferramenta de comparação, alcançaram retornos superiores aos modelos LR, e ao mercado, ao selecionar as 20 empresas com melhor desempenho e as empresas com pontuação acima de 60 %. No entanto, ao selecionar empresas com 75 % de probabilidade ou mais, o modelo LR apresentou uma capacidade fantástica de escolher as empresas certas, apresentando retornos elevados e composições de portfólio mais seguras quando comparados ao algoritmo SVM e ao índice S&P 500.

Palavras-chave: Análise Fundamental, Relatórios Financeiros, Rácios Financeiros, S&P 500, Regressão Logística, Máquina de Vetores de Suporte

Abstract

This work proposes an approach that combines Logistic Regression (LR) models with Fundamental Analysis (FA), to create a system capable of predicting long-term growing companies, when making use of the percentage scores of the algorithm. The proposed implementation fetches financial information of companies included in the S&P 500 index, from 2009 until 2021. This information is filtered and transformed into financial ratios, capable of quantifying the performance of companies, which are then used by the models. Then, the models compute probability scores that reflect the confidence level that each company will grow more than the market index, for each testing year. The companies are then ranked according to their respective probability values. The yearly returns were then evaluated by computing the Return on Investment, Max Drawdown, and Sharpe ratio, to analyze the profits generated, the volatility, and the risk associated with the portfolios computed. Over the 8 investment years, the Support Vector Machine (SVM) models, designed as a comparison tool, achieved higher returns than the LR models, and the market, when selecting the top 20 performing companies and the companies with scores above 60 %. However, when selecting companies with 75 % probability or higher, the LR model showed an uncanny ability to select the right companies, presenting high returns and safer portfolio compositions when compared to the SVM algorithm, and the S&P 500 index.

Keywords: Fundamental Analysis, Financial Statements, Financial Ratios, S&P 500, Logistic Regression, Support Vector Machines

Contents

List of Tables

List of Figures

Nomenclature

Investment Related

B&H Buy and Hold

CR Current Ratio

D/E Debt-to-Equity Ratio

DPR Dividend pay-out Ratio

DY Dividend yield

EBIT Earnings Before Interests and Taxes

EPS Earnings per share

FA Fundamental Analysis

GPM Gross Profit Margin

ICR Interest Coverage Ratio

IPO Initial Public Offering

ITR Inventory Turnover Ratio

MDD Maximum Drawdown

NI Net Income

OPM Operating Profit Margin

P/B Price-to-Book Ratio

P/E Price-to-Earnings Ratio

PEG Price/Earnings-to-Growth

QR Quick Ratio

ROA Return on Assets

ROE Return on Equity

ROI Return on Investment

S&P500 Standard and Poor´s 500 stock index

TA Technical Analysis

Optimization and Computer Engineering Related

AI Artificial Intelligence

ANN Artificial Neural Networks

ANOVA Analysis of Variance

DA Discriminant Analysis

DT Decision trees

GA Genetic Algorithm

GT Gamma Test

LDA Linear Discriminant Analysis

LR Logistic Regression

LVQ Learning Vector Quantization

MDA Multiple Discriminant Analysis

MLE Maximum Likelihood Estimation

MLR Multinomial Logistic Regression

MOEA Multi-Objective Evolutionary Algorithm

NN Neural Networks

OC1 Recursive Partitioning (Oblique Classifier)

OLS Ordinary Least Squares

PCA Principal Component Analysis

PNN Probabilistic Neural Network

QP Quadratic Programming

RI Rule Induction

SVM Support Vector Machines

Glossary

bankruptcy Bankruptcy is a legal proceeding involving a person or business that is unable to repay their outstanding debts. 24

bond rating A bond rating is a grade given to bonds that indicates their credit quality. 26

leverage Technique involving using debt (borrowed funds) rather than fresh equity in the purchase of an asset, with the expectation that the after-tax profit to equity holders from the transaction will exceed the borrowing cost. 11

liquidity Ease with which an asset, or security, can be converted into ready cash reflecting its intrinsic value. 7

profitability Measure of how much profit a company makes compared with its overall revenue and costs. 10

stock options A stock option is a contract between two parties that gives the buyer the right to buy or sell underlying stocks at a predetermined price and within a specified time period. 6

volatility Rate at which the price of a security increases or decreases for a given set of returns. In most cases, the higher the volatility, the riskier the security. 5

Chapter 1

Introduction

Stock markets are an important component in a lot of economies worldwide and play a significant role in the international financial system. It is the place where companies have their equity shares listed and investors perform trades.

For an investor, understanding what companies to invest in and when to buy and sell is key to achieving maximum performance in the stock market. However, accomplishing such a task is very difficult due to the market's dynamic and uncertain nature. In fact, the stock market is highly influenced by many different factors, from economic situations to political factors to natural calamities and many more. This is why some researchers believe in the efficient market hypothesis (Fama, 1970), stating that the market, at all times, reflects all the information available and that it is impossible to gain an advantage on it. On the other hand, others believe the opposite, that the market is inefficient and does not respond to new information quickly enough, leaving room for investors to exploit and outperform it.

Over the years, investors made use of market analysis approaches, like Fundamental Analysis (FA) and Technical Analysis (TA), in order to obtain more information on the market and on the companies. FA is more focused on finding the true value of a corporation, analyzing the financial information, and measuring performance, with a long-term view for investment. On the contrary, TA is more focused on analyzing the evolution of market prices, trying to find patterns or trends, being more valuable in short-term trading.

With the advances in Artificial Intelligence (AI) algorithms, combinations with market analysis techniques were made, with the purpose of helping investors in deciding when and where to invest. These works usually take the financial information of companies, and try to predict positive market trends or find high returning firms, depending on the strategy applied. Other researches have been done, but more focused on finding losable situations to prevent losses, like bankruptcy predictions.

Following the trend, this work proposes an implementation combining a Logistic Regression (LR) algorithm with a Fundamental Analysis approach, using the public financial data issued by companies to select the right companies to maximize the profits of investors. The hyperparameters are optimized using prior data to the testing year, using a grid search algorithm for the purpose. During the testing year, the model outputs probability scores, which are then used to rank the companies in terms of how confident the model is in their performance. Depending on the ranking, an investment simulation is made on the following year, generating results that are then analyzed. For a more complete evaluation of the

LR algorithm, a Support Vector Machine (SVM) model was also designed, in an effort to compare the proposed approach to another machine learning algorithm, and to the market, to properly validate the results and conclusions obtained.

1.1 Motivation

Investment in the stock market is an environment with many competitors and no real standout winner. Financial markets are uncertain and dynamic, so it is very hard to predict where they will be evolving next. This means there is room for improvement and for a challenge to generate better results than others. This challenge has led to many strategies being implemented using companies' financial information, market information, and economical news.

Many different pieces of research have proven that some Fundamental indicators can be of great relevance for stock movements, but markets are still affected by economical elements, and predicting stock movements can be a very hard thing to do. This leads to a need for a system that can consistently present advantageous results.

In order to find positive market trends, this research focuses on computing probabilities of stocks increasing in value, with the goal of finding long-term growth companies to obtain a higher return in the long run. For this reason, Logistic Regression is picked. Logistic Regression is not a very common algorithm in this subject of the financial world. So, FA and LR have rarely been combined to predict stock market movements. Therefore, there is a necessity to study the efficiency of this combination.

1.2 Objectives

The main goal of this work is to implement a system capable of predicting high performance companies for a year long investment in the stock market. The system is based on a Logistic Regression algorithm so that we can make use of a percentage-based method to calculate the probability of a stock value rising. This system will use a series of defined financial ratios, computed using financial information made public by each company, and will then invest in the selected companies using several investment strategies. The results will be compared with a benchmark of the market, like the S&P500 index, and an SVM algorithm designed.

1.3 Contributions

The main contributions of this thesis are:

- The combination of Logistic Regression and Fundamental Analysis to stock market prediction world. Although each topic has had its uses in several fields, the combination for long-term investment has still yet to be fully studied.
- The use of the percentage results given by the LR algorithm, computing a ranking system with the performance of the companies according to the model, to select only the safest and with higher

long-term potential companies.

- The proof given by the 16 selected features, responsible for quantifying the performance of companies, in being able to gather the necessary information to evaluate companies.

1.4 Thesis Outline

This work is organized as follows:

- Chapter 2 presents the necessary theory and methodologies used to perform this investigation including financial statements, ratios, logistic regression, and support vector machines. This chapter also presents relevant studies about the topics mentioned.
- Chapter 3 describes the architecture of the system implemented and provides an explanation to all the different layers of the system.
- Chapter 4 gives a brief description of the data and the evaluation metrics before presenting the results obtained in all test scenarios. In the end, a more detailed view of the top performing situations is made as well.
- Chapter 5 summarizes the conclusions of this work and discusses possible future work propositions.

Chapter 2

Background

This chapter will present the background and literature review of the subjects contained in this work. In section 2.1 it will be described the stock market, the place where our research is focused, and market characteristics. Section 2.2 provides a detailed explanation of Fundamental analysis, financial statements and ratio analysis. Section 2.3 will describe a few machine learning concepts, preparing for Sections 2.4 and 2.5, where both the algorithms used in this research will be explained as well as their characteristics. Section 2.7 gives insight into classification metrics that are responsible for measuring models' performance. The literature review of the topics in question is represented in section 2.8. Finally conclusions are drawn in section 2.9.

2.1 Market Analysis

The stock market is the meeting place of stock(or shares) traders. Companies make their shares available to the general public in the primary market to raise capital in Initial Public Offering (IPO). Then, in the secondary market, these sold stocks are traded between investors at the prevailing market price. This market price is influenced by the supply and demand of each listed stock and determines the ranges at which investors and traders buy and sell. These investments are mostly done through stockbrokers and are mostly based on an investment strategy. A broker is an individual or a firm that intermediates these investments since the stock exchange requires authorization to execute orders and brokers are members with that authorization. These brokers will investigate the market to invest in the best approach possible.

2.1.1 Types of Markets

Financial markets are very dynamic and uncertain, everything can change very fast. However, we can characterize the market based on the volatility of the market and the evolution of the prices. (Faith, 2007) defined four different states for markets:

- **Stable and quiet** - Prices tend to remain within a small range of values with little upward or downward variations. It is a difficult market to invest in since price movement is very low and there is no clear opportunity for profit.

- **Stable and volatile** - Prices tend to stay within a big range, with bigger value changes in small periods, but with no considerable changes in a larger period.

- **Trending and quiet** - Prices tend to evolve in only one direction, with little movement in the opposite direction over a large period.

- **Trending and volatile** - Prices tend to have larger changes in one direction with occasional significant short-term reversals of direction

2.1.2 Market Trends and Positions

The market can be described depending on its trend as bull, bear, or sideways market. A market is said to be bull when it exists an uptrend in the market, meaning prices are rising and are expected to continue rising, creating optimism for investors. On the other hand, a bear market means the opposite, with prices falling and pessimism reigning over investors. When the market is neither bull nor bear, it means there is not a clear trend, meaning prices are fluctuating inside a narrow range of values. When this happens, a market is said to be moving sideways.

Depending on the trends and types of markets, investors have to take different positions in order to maximize their positions. With this in mind, there are three possible positions: long, short, or neutral. A long position refers to the purchase of an asset believing it will increase in value in the future. A short position is established by selling a stock with plans to buy it later at a lower price. It is important to note that in short positions, the sold stock is borrowed by the broker, with the seller paying fees for borrowing the shares while the short position is active. A neutral position is taken when the market is in a sideways situation, not having a clear trend. Investors might stay out of the market, not buying or selling assets, or they can use neutral trading strategies like going short and long in similar stocks and using stock options, for example.

2.1.3 Investment Approaches

There are two investment approaches that investors use to generate profit with their transactions: active and passive investment. Active investment is a strategy that involves frequently buying and selling assets, with the desire to outperform an index or a benchmark by achieving higher returns. It is crucial to analyze market trends correctly and to understand the right moments to buy and sell. Only this way, the strategy will succeed. On the other hand, passive investment is a more conservative approach, reducing the amount of buying and selling of assets. It broadly refers to a buy-and-hold strategy for long-term investments. In the end, active investment usually produces higher returns if done correctly but with more associated risk and higher taxes and fees while passive investment is more limited but less complex.

2.2 Fundamental Analysis

With the trading business growing, and the market being as unpredictable as it is, investors started using methods in order to predict the stock market. Two major methods have been used for this purpose: fundamental analysis and technical analysis. Fundamental analysis is an approach with the goal of finding a stock's real value. Its study is based on the evaluation of several elements like financial statements, industry analysis, and ratio analysis. Financial statements consist of a company's activities and its financial records. Each one is examined to evaluate the performance of the firm. Industry analysis takes into account characteristics of an industry, like competitiveness, growth, products, risk, or customers. Ratio analysis analyses the information contained in financial statements and determines ratios that provide a quantitative value of a company's performance.

2.2.1 Industry Analysis

Industry analysis is a very useful tool for analysts since it can be a very good introduction to understanding the business where a company is inserted. Porter (1985) defined industry analysis as "Porter's Five Forces Analysis", where he distinguishes five different attributes to take into consideration based on the information that each industry has different profit margins. These attributes are industry rivalry, the threat of substitutes, bargaining power of buyers, bargaining power of suppliers, and barriers to entry. Industry rivalry refers to the competition among existing firms where an intensely competitive industry contributes to a lower profit potential for its companies. The threat of substitutes references the number of substitute products in the market where having a higher amount will restraint the possibility to raise prices. The bargaining power of buyers and suppliers reflects on the fact that powerful buyers and suppliers can reduce the profitability by playing competitors against each other, demanding more services or higher quality for example. Finally, barriers to entry considers the factors that influence a decision of a company to enter a market. These factors go from analyzing the cost advantage from already inserted firms to the portion of advertising spent by competitors being too high. These factors determine the attractiveness of a market.

2.2.2 Financial Statements

Financial statements are reports that represent the firm's activities and are issued annually or every three months. They quantify financial health, performance, and liquidity, and can provide certainty for analysts, investors, and others. Financial statements include income statement, balance sheet, and cash flow statement.

Income Statement

An income statement contributes with important insights into a firm's activity, efficiency of its operations, and performance towards the industry market. In Figure 2.1 an example of an income statement is shown. Total revenue corresponds to the amount of money made by the company due to its operations

during the period of time. Total operating expenses are the costs of keeping the business running. Operating profit, or earnings before interest and taxes (EBIT), stands for the profits a company makes from its core activities. Net income (NI) refers to the amount of revenue remaining after deducting expenses, taxes, and costs as shown in (2.1):

$$\text{Net Income} = (\text{Total Revenue} + \text{Gains}) - (\text{Total Expenses} + \text{Losses}) - \text{taxes} - \text{interest} \quad , \qquad (2.1)$$

with gains being income from non-business operations and non-operating losses being losses on unusual costs like assets or lawsuits.

Revenues	€
Cash sales	91,196
Credit sales	43,273
Total Revenue	**134 469**
Cost of goods sold	62,876
Gross Profit	**71,593**
Operating expenses	
Salaries	23,189
Advertising	6,240
Office rent	7,250
Utilities	4,400
Office supplies	650
Depreciation	4,240
Other expenses	3,300
Total Operating expenses	**49,269**
Operating profit	**22,324**
Operating income	
Interest income	600
Interest expenses	2,200
Net income before tex	**19,524**
Income tax expenses	3,100
Net income after tax	**16,424**

Figure 2.1: Example of an Income Statement.

Balance Sheet

A balance sheet provides information about what a company owns, owes and how it is financed. It is divided in three categories: assets, liabilities and shareholders' equity, which is shown in Figure 2.2. Their relationship is demonstrated in (2.2),

$$\text{Assets} = \text{Liabilities} + \text{Shareholders' Equity} \quad . \qquad (2.2)$$

Current Assets	€
Cash and cash equivalents	14,196
Receivables	5,273
Allowance for Doubtful Accounts	(65)
Inventories	10,876
Other current assets	1,866
	32,146
Non-current Assets	
Equipment	64,834
Vehicles	95,524
Accumulated depreciation	(31,098)
	127,260
Total Assets	159,406
Current liabilities	
Interest payables	1,865
Accounts payables	6,599
Accruals	3,100
Other current liabilities	3,987
	15,551
Non-current liabilities	
Note payables	39,844
	39,844
Total Liabilities	55,395
Equity	
Share capital	58,136
Retained earnings	31,876
Profit/(Loss) current year	13,999
Total Equity	104,011
Total liabilities and Equity	159,406

Figure 2.2: Example of a Balance Sheet.

Assets are registered in terms of liquidity i.e., in the beginning, the ones that are easier to turn into cash are displayed. They are divided into two parts: current assets and non-current assets. Current assets are the assets that can be converted to cash in less than a year while non-current assets are the assets that cannot.

Liabilities correspond to the money a company owes to third parties. These include payments to suppliers, interest on bonds, wages, and short-term debt. Like assets, liabilities are also divided into two categories, current liabilities and non-current liabilities, which correspond to those that are due in less than a year and those that are due in over a year, respectively.

Shareholders' equity is the capital attributable to the company's owners. It usually includes retained earnings, common stock, preferred stock, and treasury stock. Retained earnings are the amount of NI the company uses to pay a debt or reinvest in the business. The remaining is divided through sharehold-

ers. Common stock is a financial asset that represents ownership of a company, provides the power to elect a board of directors and to earn dividends. Preferred stock is also a financial asset that only gives the right to collect dividends before common stockholders receive theirs. Treasury stock corresponds to the stock a firm has bought back.

Cash Flow Statement

A cash flow statement sums up the amount of money getting in and getting out of a company. It estimates how good of a position a company is in by evaluating if a firm is generating enough cash to pay its debt and to back its operations. As shown in Figure 2.3, this is divided into three categories: cash from operating activities, cash from investing activities, and cash from financing activities. Cash from operating activities indicates how much money is made by a company's business. These may include interest payments, employees' salaries, rent, and payments to suppliers, amongst others. Cash from investing activities shows the cash movement related to the company's investments. These investments include asset purchase, equipment purchase, loans made or received, amongst others. Cash from financing activities displays the cash flows that fund the company. These include loans from banks or investors, dividends, stock repurchases, or debt payments. These cash flows are positive when money is coming into the company and are negative when money is flying out of the company.

Cash flow from operations	€
Net income	81,000
Adjustments for depreciation	3,000
Adjustments for increase in inventories	(26,000)
Adjustments for decrease in accounts receivable	13,000
Net cash flow from operations	**71,000**
Cash flow from investing	
Cash receipts from sale of property and equipmen	9,000
Cash paid for purchase of equipment	(13,000)
Net cash flow from investing	**(4,000)**
Cash flow from financing	
Cash paid for loan repayment	(5,500)
Net cash flow from financing	**(5,500)**
Net increase in Cash	**61,500**

Figure 2.3: Example of a Cash flow statement.

2.2.3 Ratio Analysis

Ratio analysis is a quantitative method to investigate a company's profitability, liquidity, and business efficiency by looking at its financial statements. It compares current years' ratios with the ones from

previous cycles, to evaluate its evolution and compute its performance. Some of the ratios generally used by investors are listed below.

- **Quick Ratio (QR)**

  The quick ratio is a pointer of a company's short-term liquidity position and measures a company's capability to satisfy its short-term obligations with its near-cash assets. Typically, a value bigger or equal than one indicates a company that can meet its obligations where the higher the number, the better the situation of the company.

$$\text{Quick Ratio} = \frac{\text{Cash \& equivalents + marketable securities + accounts receivable}}{\text{Current Liabilities}} \tag{2.3}$$

- **Current Ratio (CR)**

  This ratio measure the ability of the company in covering its short-term obligations with short-term assets. A higher figure is a sign of fiscal strength while a value lower than one is a matter of concern.

$$\text{Current Ratio} = \frac{\text{Current Assets}}{\text{Current Liabilities}} \tag{2.4}$$

- **Debt-to-Equity Ratio (D/E)**

  The D/E ratio helps investors determine how a company finances its assets. It shows how much debt is engaged in the corporation. Generally, a low value of D/E indicates a lower risk for shareholders but it is a ratio that should be looked at amongst other ratios and compared across the same industry.

$$\text{Debt-to-equity Ratio} = \frac{\text{Total Liabilities}}{\text{Shareholders' Equity}} \tag{2.5}$$

- **Interest Coverage Ratio (ICR)**

  This leverage ratio is used to figure out how handily a company can pay interest on its outstanding debt. A higher value is generally better but the industry may influence the conclusions that can be taken.

$$\text{Interest Coverage Ratio} = \frac{\text{EBIT}}{\text{Interest Expense}} \tag{2.6}$$

- **Asset Turnover Ratio (ATR)**

  This ratio measures the efficiency with which a company uses its assets to produce sales. A higher value indicates that a company is more efficient than its competitors. Like many other ratios, investors use it to study companies in the same industry.

$$\text{Asset Turnover Ratio} = \frac{\text{Total Revenue}}{\text{Average Assets}} \tag{2.7}$$

- **Inventory Turnover Ratio (ITR)**

  Inventory turnover shows the number of times a company has bought and sold inventory over some time. A high ratio value indicates substantial sales or insufficient inventory while a low value can indicate frail sales or excess inventory. It should be used in comparison with companies in the same industry since benchmark values depend on it.

$$\text{Inventory Turnover Ratio} = \frac{\text{Cost of Goods Sold (COGS)}}{\text{Average Inventory}} \tag{2.8}$$

- **Return on Equity (ROE)**

  ROE measures how profitable a company is in relation to stockholders' equity. Although a high value is usually positive which would mean that a company can generate good income on new investment, other factors can inflate the ROE value. In conclusion, it is very difficult to retrieve definite conclusions only using this ratio.

$$\text{Return on Equity} = \frac{\text{Net Income}}{\text{Shareholders' Equity}} \tag{2.9}$$

- **Return on Assets (ROA)**

  Return on assets is an indicator that shows the percentage of profit a company generates relative to its total assets. ROA is best used when compared to past figures of this indicator since it varies significantly from industry to industry. In this case, a positive trend over the years is a good sign for the company where the greater the ROA, the more money a company is making on less financing.

$$\text{Return on assets} = \frac{\text{Net Income}}{\text{Total Assets}} \tag{2.10}$$

- **Gross Profit Margin (GPM)**

  Gross profit margin is used to estimate how much profit a company makes from its sales after deducting the cost of goods sold. It is mostly utilized to compare companies inside an industry.

$$\text{Gross profit margin} = \frac{\text{Total revenue - COGS}}{\text{Total Revenue}} \tag{2.11}$$

- **Operating Profit Margin (OPM)**

  Operating margin measures how much profit a company generates from its main operations. A good way of looking at it is by comparing its past figures and seeing the evolution. A positive tendency over a while means the profitability of a company is increasing.

$$\text{Operating Profit Margin} = \frac{\text{EBIT}}{\text{Total Revenue}} \tag{2.12}$$

- **Price-to-Earnings Ratio (P/E)**

The P/E ratio is the ratio for valuing a company that estimates its current share price relative to its earnings per share (EPS). It is usually used to determine if a company is being overvalued or undervalued. A lower P/E ratio might indicate a good opportunity for investors.

$$\text{Price-to-Earnings Ratio} = \frac{\text{Market Value per Share}}{\text{Earnings per Share}} \tag{2.13}$$

- **Price/Earnings-to-Growth (PEG)**

  The PEG ratio is an indicator used to measure a company's stock value. It is interrelated with the P/E ratio where PEG is used to confirm the results obtained in the P/E ratio. A value of one means the stock is fairly priced while a value below one may illustrate that a stock is undervalued.

$$\text{Price/Earnings-to-Growth} = \frac{\text{Price-to-Earnings}}{\text{Earnings per Share Growth Rate}} \tag{2.14}$$

- **Price-to-Book Ratio (P/B)**

  The P/B ratio measures the market's valuation of a company relative to its book value. Book value corresponds to the amount of money that would result if a company would sell all assets and repay all liabilities. A value of P/E below one might indicate an undervalued stock.

$$\text{Price-to-Book Ratio} = \frac{\text{Market Value per Share}}{\text{Book Value per Share}} \tag{2.15}$$

- **Earnings per Share (EPS)**

  EPS indicates how much money a company makes per share of common stock. It is used by investors to measure a company's profitability where, the greater the value, the more lucrative the company is regarded. EPS should be visualized with its past values and compared to similar companies.

$$\text{Earnings per Share} = \frac{\text{Net Income - Preferred Dividends}}{\text{Average Outstanding Shares}} \tag{2.16}$$

- **Dividend Yield (DY)**

  It is a financial ratio that shows the percentage of return through dividends relative to a company's stock price. A lot of factors can influence the value of dividend yield i.e., it is not a good indicator when looking for a growing company. However, investors use this ratio when looking for companies that pay high dividends to their shareholders.

$$\text{Dividend Yield} = \frac{\text{Dividend per Share}}{\text{Market Price per Share}} \tag{2.17}$$

- **Dividend Pay-Out Ratio (DPR)**

  Dividend Pay-Out measures the portion of earnings that is allocated to shareholders. A high value of DPR means the company is paying most of its net income as dividends, while a low value means

the company is reinvesting most of its earnings in the company to expand and grow.

$$\text{Dividend Pay-Out} = \frac{\text{Dividend per Share}}{\text{Earnings per Share}} \tag{2.18}$$

2.3 Machine Learning Concepts

2.3.1 Supervised and Unsupervised Learning

There are two methods for training machine learning algorithms: supervised and unsupervised learning. Supervised learning, which is used in this work, consists of the models training with labeled data. This means that the models try and find a mapping function that maps the input variable to its output. Since the output is known, the model receives feedback, checking if the predictions are being correct or not. This method is mostly used for classification or regression problems. Unsupervised learning is used to find patterns and the structure of the data. This is because there is no output known for the data which means that the algorithm can not map the relations. Having no output in the data means that the model has no feedback. This method is mostly used in clustering and associations problems.

2.3.2 Overfitting and Underfitting

When training an algorithm, there are two main motives for poor performance of the models: overfitting and underfitting. Overfitting happens when a model learns the training data too well, leading to it understanding the detail and noise of the data. The problem is that these concepts do not affect other datasets the same, ending up influencing the negative generalization ability of the model. This is usually the case when an algorithm shows great results in the training set but has significantly lower ones in the validation or testing sets. Underfitting refers to a model that can not perform in the training data nor obtain good results in new data provided. Out of the two, overfitting is the most common problem in machine learning algorithms. A representation of these two situations is shown in Figure 2.4

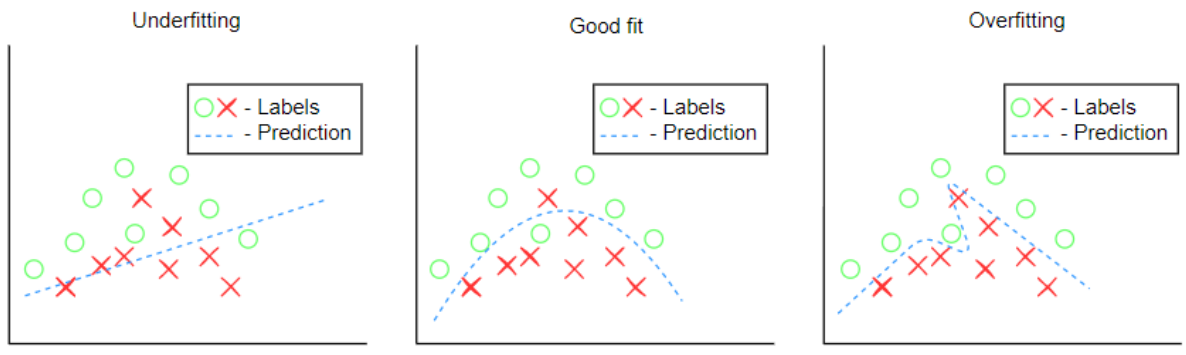


Figure 2.4: Overfitting and Underfitting.

2.3.3 Cross-Validation

Cross-Validation is a resampling program to evaluate machine learning algorithms with a limited amount of data. There is only one parameter, K, that designates the number of groups, or folders, that a dataset will be split into. This procedure is applied to the training data to validate the solution obtained so that it can be evaluated on the testing set. An example of Cross-Validation is represented in Figure 2.5.

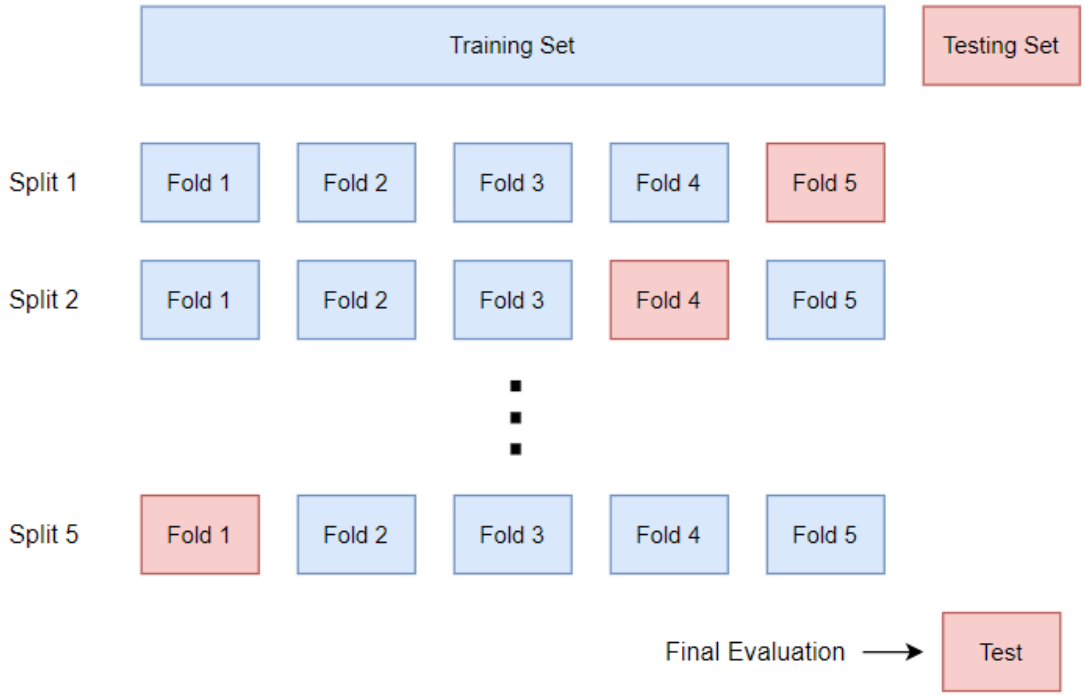


Figure 2.5: K-Fold Cross-Validation.

The procedure is very simple. Using a K of 5, for example, the data is split into 5 folders of the same size, and then, for each iteration, the model is trained with the data of every folder except one, which will be used for testing. Ultimately, the method will return K metric values, to properly evaluate the abilities of the model.

2.4 Logistic Regression

Logistic Regression is a statistical method mostly used to inspect a relationship between a dependent variable and one or more independent others. Generally, it is used for classification purposes where the interest is to figure out the probability of an event happening. LR computes a sigmoid function that takes a real input and determines an output between zero and one. This function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{\ z}} \quad , \tag{2.19}$$

where z being a multi variable linear function which is shown in equation (2.20),

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n = X^T \beta \quad . \tag{2.20}$$

Where β_i (i= 0, 1, ..., n) correspond to the weights assigned to the input variables and x_i (i= 0, 1, ..., n) being the input independent variables.

In the end, the logistic function can be represented by:

$$p_i = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \quad , \tag{2.21}$$

with $p_i = P(Y_i = 1)$ representing the estimated probability of a situation happening.

One way of quantifying the relation between output and input is with odds ratios. Odds ratio correspond to the probability of an event occurring divided by the probability of it not occurring. In this research, the event corresponds to the value of a stock increasing. We can obtain the odds ratio from equation (2.21):

$$=) \quad 1 \quad p_i = 1 \quad \frac{1}{1 + e^{z}} = \frac{1 + e^{z} \quad 1}{1 + e^{z}}$$

$$= \frac{e^{z}}{1 + e^{z}} \quad =) \quad \frac{p_i}{1 \quad p_i} = \frac{1}{e^{z}} = \text{Odds} \quad .$$

Taking the natural logarithm of both sides:

$$ln\left[\frac{p_i}{1 \quad p_i}\right] = \quad (\quad z) = z = X^T\beta \quad . \tag{2.22}$$

The coefficients β_i (i= 0, 1, ..., n) will be estimated using the sklearn python library, by importing the Logistic Regression model and utilizing all of its tools. This model has five different solvers available to solve the optimization problem.

2.4.1 Solvers

The five solvers available in the Logistic Regression model are: Newton-cg, Lbfgs, Liblinear, Sag, and Saga. Each one has a different way to solve the optimization problem and for this reason, each is more suited for different situations.

- **Newton-cg**

  Newton's method consists of a quadratic function minimization problem. It uses the Hessian matrix as the quadratic function, which consists of the first and second partial derivatives as can be seen in Figure 2.6.

$$\mathbf{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Figure 2.6: Definition of a Hessian Matrix.

Basically, at each iteration, $f(x)$ is approximated by a quadratic function around x, and then takes a step in the direction of a maximum or minimum of the function.

The problems with Newton's method are the computational cost, due to the use of the Hessian matrix and the second partial derivative calculations, and the fact that it attracts saddle points (points where the partial derivatives disagree on whether it is a maximum or a minimum point) in higher dimension problems.

- **Lbfgs**

  The Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm, or Lbfgs, is a similar method to Newton's method explained before but with a critical difference, it doesn't calculate the exact Hessian matrix but calculates approximate gradient evaluations instead, making it computationally more effective. The term Limited-memory references the fact that it only keeps the latter gradient evaluations, discarding the initial ones.

  In small datasets, the Lbfgs has proven to perform well compared to other solvers, saving a lot of memory due to its properties.

- **Liblinear**

  It's a linear classifier that supports logistic regression and linear support vector machines. A linear classifier makes a decision based on the result of a linear combination of the features.

  The solver uses a coordinate descent (CD) algorithm that resolves optimization problems by, at each iteration, operating approximate minimization along coordinate directions.

  Liblinear solver is more efficient than others for small datasets but has some drawbacks. It will have a hard time with multiclass problems as it separates the problems into a "one-vs-rest" situation and it only supports l1 penalization.

- **Sag**

  The Sag, or Stochastic Average Gradient, optimizes the sum of various convex functions. It is similar to other stochastic gradients (SG) in terms of the cost of iterations being independent of the

number of terms in the sum. On the other hand, the Sag method can reach faster convergence rates than other SG methods since it incorporates a memory to store past gradient values.

Unlike Lbfgs and Liblinear, this solver is more adequate for large datasets, being faster than other solvers. However, it is also limited as it only supports l2 penalization.

- **Saga**

  The Saga method is very similar to the Sag method described but also supports l1 and elasticnet penalization. In the end, it ends up being the most suited solver for very large datasets and sparse multinomial logistic regression.

2.4.2 Advantages vs Disadvantages

Logistic Regression is a very simple method to implement, train and decipher. Being a very easy technique to implement and having a training time far more reduced than other more complex methods, LR is considered a good benchmark to estimate performance. It is a model that does not make assumptions of variable distributions and presents an estimate of how important a variable is, both positively or negatively. It is concluded to be a very appropriate method to implement when data is linearly separable since it shows good accuracy and efficiency.

However, Logistic Regression is not applicable in many cases. Linearly separable data is not very common in real problems which means LR cannot be applied. It is a model that has a hard time obtaining complex relations where other algorithms thrive. It is very sensitive to outliers, which means that it might provide incorrect results due to their presence in the data.

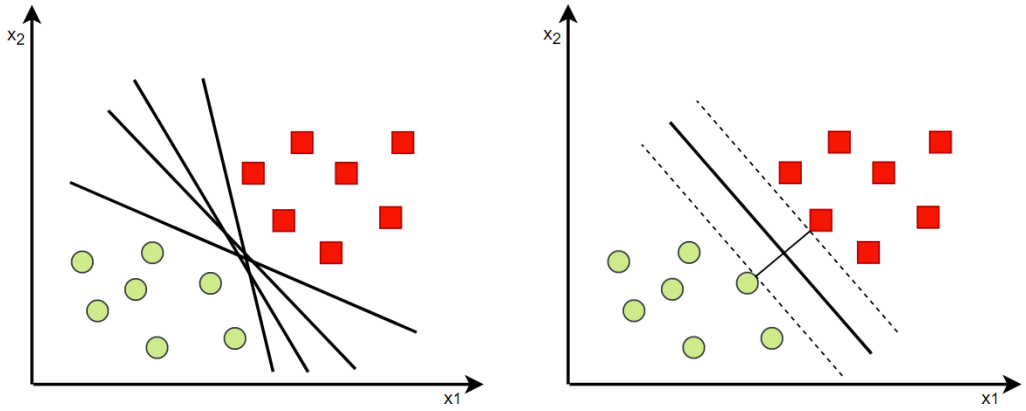
In conclusion, Logistic Regression is a technique more suitable to solve simpler problems where linearly separable data is abundant. However, such problems are rarely found in the real world which leads to an under-use of this model.

2.5 Support Vector Machine

Support Vector Machines are supervised learning algorithms that utilize classification methods, based on decision boundaries, for two-group classification problems. This approach was advanced by Vladimir Vapnik (Vapnik and Lerner, 1963) and is, nowadays, one of the most used Machine Learning algorithms due to its simplicity and high performance even with a limited amount of data (Pisner and Schnyer, 2020).

The main intention of an SVM is to find the best hyperplane in an N-dimensional space, with N being the number of features, that can distinguish data points according to their classification. For any dataset, there will be multiple hyperplanes that will perform this classification correctly, with an even higher number of possible solutions if the dataset has a limited amount of samples. So, from all these options, the best solution will be the hyperplane that maximizes the distance between the data points closest to the decision boundary (support vectors). This is called ”Widest Street Approach”. Different support vectors will result in different hyperplanes being computed. For example, in Figure 2.7(a), data

points of each class are given as well as multiple hyperplanes that would satisfy the problem at hand. In Figure 2.7(b), the best solution is given, with the hyperplane that maximizes the distance between the support vectors (margin). In the end, the bigger the margin, the bigger the confidence in correctly classifying future data points.



(a) Multiple Solutions (b) Best Solution

Figure 2.7: Widest Street Approach.

2.5.1 Widest Street Approach

Assuming samples $x_1, x_2, ..., x_m$ in a feature space H, the hyperplanes in H can be written as

$$f\vec{x}\ 2\ Hjh\vec{w}, \vec{x}i + b = 0g, \vec{w}\ 2\ H, b\ 2\ \mathbb{R} \tag{2.23}$$

where $\vec{w}$ corresponds to the vector orthogonal to the hyperplane and $\vec{x}$ to a new sample vector. Equation 2.23 comes from the desire to classify a vector $\vec{x}$ as belonging on one side of the hyperplane. For that, vector $\vec{x}$ is projected on vector $\vec{w}$, as an internal product $\vec{w}\ \ \vec{x}$. This will return the length of $\vec{x}$ in the direction of $\vec{w}$ and, if this value is either bigger or smaller than the value of a constant C, then the sample will be classified accordingly to the side of the hyperplane it belongs. In the end, the decision rule for a sample whose distance is bigger than C is given by $\vec{w}\ \ \vec{x} + b\ \ \ 0$, with $C = \ \ b$. Vapnik ended up proposing two constraints, one for each side of the decision boundary, corresponding to $\vec{w}\ \ \vec{x} + b\ \ \ 1$ and $\vec{w}\ \ \vec{x} + b\ \ \ 1$ which can be written as $y_i(\vec{w}\ \ \vec{x} + b)\ \ \ 1$, where y_i will have a 1 or $\ \ 1$ value depending on the side of the hyperplane.

Given this, to find the best solution, one needs to maximize the distance between support vectors, which would be $\frac{2}{k\vec{w}k}$. Maximizing this expression is equivalent to minimizing $k\vec{w}k$, or $k\vec{w}k^2$ since $k\vec{w}k^2$ is an increasing function. So, for mathematical convenience, the expression $\frac{1}{2}k\vec{w}k^2$ will be minimized leading to the problem described in 2.24.

$$\begin{aligned} &\min_{\vec{w}2H,b2\mathbb{R}} \quad \frac{1}{2}k\vec{w}k^2 \\ &\text{subject to} \quad y_i(\vec{w}\ \ \vec{x_i} + b)\ \ \ 1 \quad 8i = 1, ..., m \end{aligned} \tag{2.24}$$

First, and with the use of the method of Lagrange multipliers $\alpha_m\ 2\ \mathbb{R}$, the Lagrangian is computed

and the partial derivatives are estimated as shown in 2.25 and 2.26, respectively.

$$L = \frac{1}{2}k\vec{w}k^2 \quad \sum_{i=1}^{m} \alpha_i(y_i(\vec{w} \quad \vec{x_i} + b) \quad 1) \tag{2.25}$$

$$\begin{aligned} \frac{\partial L}{\partial w} = w \quad \sum_{i=1}^{m} \alpha_i y_i x_i = 0, \\ \frac{\partial L}{\partial b} = \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned} \tag{2.26}$$

Second, the derivatives are replaced in the Lagrangian function, resulting in

$$L = \sum_{i=1}^{m} \alpha_i \quad \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \quad x_j) \tag{2.27}$$

By observing 2.27, it is possible to conclude that the optimization of the hyperplane is only influenced by the support vectors and that the equations that rule the decision boundary are

$$\begin{aligned} \sum_{i=1}^{m} \alpha_i y_i x_i \quad \vec{x} + b \quad 0 \\ \sum_{i=1}^{m} \alpha_i y_i x_i \quad \vec{x} + b \quad 0 \end{aligned} \tag{2.28}$$

2.5.2 Kernels

The descriptions given before characterise the simple example of separating samples whether it is with a straight line or in a higher dimension space that can not be visualized. Like it was said before, the decision rule is only influenced by the support vectors, which means the dimension of the vectors does not have a saying. The problem is when the samples can not be separated by a straight line. To solve this situation, Vapnik introduced what is called the ”kernel trick” (Boser et al., 1992), a transformation function $\psi(\vec{x})$, that will bring the samples to higher dimension spaces until they can be separated. Basically turning a nonlinear classification rule into a linear rule of the transformed sample points. Now, based on what has already been explained before, to maximize the distance between the support vectors, the inner product is done $\psi(\vec{x_i}) \quad \psi(\vec{x_j})$. According to Mercer's Theorem, this inner product can be defined through a kernel $K(x_i, x_j)$, being this kernel that will maximize the margin.

The kernels for support vector classification tested in this work are presented in table 2.1.

Table 2.1: SVM kernels.

Kernels	Functions
Linear	$K(x_i, x_j) = x_i^T x_j + c$
n order Polynomial	$K(x_i, x_j) = (x_i^T x_j + c)^n$
Radial Basis Function	$K(x_i, x_j) = \exp \quad \frac{kx_i \quad x_j k}{\sigma}$
Sigmoid	$K(x_i, x_j) = tanh(\alpha x_i x_j + c)$

2.5.3 Platt Scaling

Support Vector Machines are algorithms that function purely as a classifier, outputting scores that define the final result. So, to obtain a probability distribution over classes, one has to use Platt Scaling. Platt Scaling is a method invented by John Platt when using SVMs to address this very problem (Platt, 2000).

Platt Scaling uses a logistic transformation, as shown in 2.29, of the scores obtained by the classifier to produce the probability estimates.

$$P(Y = 1|x) = \frac{1}{1 + e^{Af(x)+B}} \quad , \tag{2.29}$$

where f(x) are the classifier's scores and A and B are two parameters estimated by the algorithm, by using a maximum likelihood method optimized with the same training set as the classifier.

Platt Scaling has proven to be impressive when used for different classification models as SVMs or naive Bayes classifiers, as these methods produce distorted probability distributions. An alternative to Platt Scaling is Isotonic Regression, which has proved to work better than the previous method, but only when the available data is sufficient since it tends to overfit.

2.6 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction procedure that is used to identify a small number of uncorrelated features, called principal components, from a large set of variables. The goal of PCA is to use the least number of principal components to explain the maximum amount of variance, resulting in a smaller portion of the data while keeping as much information as possible.

Analyzing data in high dimensional spaces can have problems related to the curse of dimensionality (Bellman, 1952), which states that, when the number of dimensions rises, the volume of the space increases so quickly that the data becomes sparse. With the sparsity of the data, it is very difficult to reach a reliable result and, to obtain such a result, the amount of data required grows exponentially with dimensionality. High dimensionality data can also have problems related to a high computational cost or overfitting, which ultimately will result in a worse performance of the model.

PCA converts the data to principal components. These principal components are linear combinations of the features in the data set. The number of principal components is the same as the number of features in the data. Each one is obtained as shown in 2.30, where X is the variables of the data set, Y_i is the principal component, $\vec{\alpha_i}$ is the coefficient vector and n is the number of features in the data.

$$Y_i = X\alpha_i = \begin{bmatrix} \alpha_{1,1}.X_1 + \alpha_{1,2}.X_2 + \alpha_{1,3}.X_3 + ... + \alpha_{1,n}.X_n \\ \alpha_{2,1}.X_1 + \alpha_{2,2}.X_2 + \alpha_{2,3}.X_3 + ... + \alpha_{2,n}.X_n \\ \alpha_{3,1}.X_1 + \alpha_{3,2}.X_2 + \alpha_{3,3}.X_3 + ... + \alpha_{3,n}.X_n \\ \vdots \\ \alpha_{n,1}.X_1 + \alpha_{n,2}.X_2 + \alpha_{n,3}.X_3 + ... + \alpha_{n,n}.X_n \end{bmatrix} \tag{2.30}$$

The coefficient vector $\vec{\alpha_i}$ is estimated by solving 2.31, where the goal is to maximize the variance of

each principal component, with the constraint that, each linear combination, has to be orthogonal to the preceding others so that each principal component picks up new pieces of information.

$$\begin{aligned} \text{Max} \quad & \vec{\alpha_i} = Var(Y_i) \\ \text{subject to} \quad & \vec{\alpha_i}^T.\vec{\alpha_i} = 1 \\ & Cov(Y_i, Y_j) = \vec{\alpha_i}^T.\lambda_i.\vec{\alpha_i}, \quad j = 1, 2, ..., i \quad 1 \end{aligned} \tag{2.31}$$

with $\lambda_i = Cov(\vec{\alpha_i}^T X)$. A more in depth description of the PCA procedure is explained in (Jolliffe, 2002).

In the end, PCA reduces the dimensionality by keeping only a defined number of principal components, as long as these pick up most of the information.

2.7 Classification Metrics

Having five different solvers to test and various parameters to tune, there is a need to measure the performance of each model to pick the best one for each case. With this in mind, classification metrics quantify the performance in different ways to have a more precise view of each model.

- **Accuracy**

  Accuracy, presented in 2.32, is probably the most used classification metric. It measures the fraction of correct predictions in a dataset.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2.32}$$

  As one can easily understand from 2.32, a higher accuracy value means that the model is correctly predicting well. However, accuracy can be misleading. When working with class imbalanced datasets, which means a high disparity between positive and negative labels, accuracy might often not tell the whole story. For example, in a dataset with 99 class 1 values and 1 class 0 values, a model might predict 98 correct class 1 values and 1 correct class 0 value, scoring a 99% accuracy. On the other hand, that is the same score as a model that always predicts class 1 values, having no predictability ability to distinguish classes. In the end, accuracy is a very useful metric to measure performance, but with the help of others.

- **Precision and Recall**

  Precision, shown in 2.33, is the ratio of correct positives to all the positive predictions by the model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{2.33}$$

  Recall, given by 2.34, is the ratio of true positives to all the positives in the dataset.

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{2.34}$$

Precision and recall are two metrics that are usually used together. However, the two are often in opposing trends, that is, increasing precision typically means reducing recall and vice-versa.

- **F1-score**

  F1-score, presented in 2.35, is calculated from the precision and recall values and used when the goal is to find an optimal balance of the two.

$$\text{F1-Score} = 2 \quad \frac{\text{precision} \quad \text{recall}}{\text{precision} + \text{recall}} \tag{2.35}$$

  An f1-score of 1 is considered perfect and means that there is a low amount of false positives and false negatives in the predictions, while an f1-score of 0 just means the model is a complete failure.

- **ROC-AUC**

  A ROC curve, or receiver operating characteristic curve, is a graph that shows the performance of the model at the probability thresholds. It plots the true positive rate (TPR), also known as recall, versus the false positive rate (FPR). These two parameters are presented in 2.36 and 2.37.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.36}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.37}$$

  To evaluate the points of the ROC curve, we use the AUC, which stands for area under the curve, which measures the area underneath the entire line. This way, it is possible to evaluate the performance across multiple probability thresholds with an AUC value of 1, representing a model with 100% correct predictions, and 0 representing a model with 0% true ones.

- **Confusion Matrix**

  The confusion matrix is a table that supports the visualization of the performance of a model. It presents a more perceptive view of which classes are being well predicted and also the type of errors that are being made. The full guide to the confusion matrix can be seen in Figure 2.8.

		True Value	
		Positive	Negative
Predicted Value	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

Figure 2.8: Guide to the confusion matrix.

2.8 Related Work

2.8.1 Fundamental Analysis

The economy of a country, or even the world, can be vastly impacted by a financial crisis. The collapse of a bank or an important company can bring significant repercussions e.g., people lose their jobs, creditors lose their loaned money and investors lose their corresponding stocks. Additionally, it can influence the market itself and cause problems for other companies. Therefore, this is a subject that has gained attention over the years and triggered a search for a model that could predict such episodes.

(Altman, 1968), one of the first in the building of methods for risk management and bankruptcy forecasting, developed a formula that predicts the possibility of a company collapsing within two years. With this goal in mind, he used a set of financial ratios with a Multiple Discriminant Analysis (MDA) approach.

Another subject where FA is generally used is in stock market performance prediction. For an investor, being able to evaluate the potential of a play and foresee the market's tendency is a huge weapon in a very competitive environment. However, this accomplishment is not a very easy task. Past data can be unreliable due to the existence of non-linearity and characteristics of each different industry. (Mubin et al., 2014) tested the efficiency and importance of different ratios amongst a variety of different industries with an Analysis of Variance (ANOVA) method and regression analysis. This gives the idea that the timeline of the data and the industries chosen play an important role when trying to develop this model.

In the early '80s, (Basu, 1983) tested the relationship between earnings' yield, firm size, and returns on the common stock and reached the conclusion that the common stock of companies with higher price-to-earnings value, earn bigger risk-adjusted returns than the common stock of firms with lower price-to-earnings value. The use of specific ratios to analyze stock market performance is not rare, with different researches looking to simplify the problem at hand, and finding the correct ratios that capture the essence of the stock. (Fama and French, 1992) designed the three-factor model which combined firm size, book-to-market equity, and excess return on the market. In (Fama and French, 2015), (Fama and French, 2016) and (Fama and French, 2017), the model was then changed to five-factor adding profitability and investment factors to the ones mentioned before.

(Greenblatt, 2005) wrote a book where he developed an investment strategy based on 2 fundamental indicators and gave it the name "Magic Formula". It uses price-to-earnings, or EV/EBIT, to identify how cheap a stock is and the return on invested capital to measure the quality of the company, with the aim of investing in quality firms at a cheap price. However, (Heegaard and Sørensen, 2013) investigated the "Magic Formula" and concluded that, although it can provide valuable observations as to which stocks to choose from, its results can be affected very significantly due to the presence of certain stocks that would differ from the final values. They added that this model should only be considered as a starting point instead of an actual explicit measure.

(Olson and Mossman, 2003) used financial ratios to compare Neural Networks (NN) with LR and Ordinary Least Squares (OLS) in forecasting Canadian stock returns, and concluded that NN outperforms the others. It was also highlighted that the choice of the ratios and the specifications of the filter's cut-off may represent a variable as important as the choice of investment techniques. (Quah, 2006) mixed financial ratios, from different financial categories, with Artificial Neural Networks (ANN) to beat the performance of the market and reported encouraging results defending that ANN has better generalization capacity than conventional methods.

(Albanis and Batchelor, 2007) developed a new approach where it was utilized fundamental ratios to test the viability of combining five different statistical classifiers (Linear Discriminant Analysis (LDA), Learning Vector Quantization (LVQ), Probabilistic Neural Network (PNN), Recursive Partitioning (Oblique Classifier OC1), Rule Induction (RI)). Results showed that combining the statistical methods through the unanimous voting principle significantly improves results in regards to the majority voting rule. (Ali et al., 2018) also attempted to predict stock performance using a combination of accounting and financial ratios, in the Pakistan Stock Exchange (PSE). Results obtained indicate that return on equity, earnings per share, sales growth, price-book ratio, current ratio, and debt to equity can be used as an indicator of a firm's performance.

A still unanswered question in the financial world is which analysis is more effective, if fundamental or technical. However, a combination of both strategies can provide a better glimpse of the market and what the best opportunities consist of. (Silva et al., 2015) combined both analysis methods mentioned and a Multi-Objective Evolutionary Algorithm (MOEA) intending to compose a portfolio that would outperform the market. Not only does this research show promising results by returning above-average returns but by also concluding that the increment of fundamental indicators can provide a boost on results and improve the precision of the simulations.

(Tsai et al., 2011) combined economical, financial, and technical ratios to test the performance of classifier ensembles and single classifiers (NN, Decision trees (DT), and LR). Their results showed that classifier ensembles outperformed single classifiers in the areas of return on investment and accuracy prediction.

The work shown above proves the importance of Fundamental Analysis in the investment world. Inside the financial subject, it has shown promising results in different researches, always by analyzing the true value of companies.

2.8.2 Logistic Regression

Logistic Regression is a model mostly used with the purpose of a classifier than as a method for forecasting out-performing stocks. Inside the financial market, LR has been mostly used in an environment of corporate finance like the prediction of corporate bankruptcy or financial ill firms.

(Ohlson, 1980) made use of LR in his research to predict collapsing or financially distressing companies. He made two particular conclusions: that any model is highly dependant on the information available and that the model designed shows strong predictive capabilities in the matter in hands. (Zavgren, 1985) developed a Logistic Regression model to also evaluate bankruptcy in firms to detect signs of ailing companies up to five years before their collapse. They concluded that the model designed showed very significant results compared to a discriminant analysis method.

(Min and Jeong, 2009) used a Genetic Algorithm (GA) in comparison to other classification models like DT, Discriminant Analysis (DA), NN, and LR, to predict bankruptcy or non-bankruptcy firms. Results showed LR with an accuracy of about 71 %, while other methods like Decision Trees and Neural Networks got a higher 76 % rate. Adopting a multivariate logit model, (Gunsel, 2005) tested the effectiveness of a series of ratios in predicting bank collapse. He obtained that low capital requisite, high leverage, big interest expense, low profitability, liquidity ratio, and small asset size have a big connection to the possible result of a bank collapse.

Also to forecast corporate bankruptcy, (Chen, 2011) mixed financial and non-financial ratios with Decision Trees and Logistic Regression methods. The model mixed the algorithms mentioned with Principal Component Analysis (PCA) to find the most suitable fundamental factors. He discovered that the PCA had a negative impact on the DT classification and very little impact on the LR model. He also concluded that DT has a greater accuracy value in situations where the collapse of the firm was in a near future, period lower than one year, while LR demonstrates better predictions in a longer run.

As a classifier, LR is not only explored in the context of financial distress. (Öğüt et al., 2009) used Artificial Neural Networks and Support Vector Machines (SVM) to detect stock manipulations in the Istanbul stock exchange. They compared their performance with other statistical methods like Logistic Regression and reached the conclusion that ANN and SVM out-perform LR in performance and classification of manipulated instances while LR showed better results in non-manipulated occurrences. (Maher and Sen, 1997) utilized an LR model to compare with a Neural Network approach to predict bond rating. He discovered that NN obtain better results when compared to a more conventional method like LR since the developed approach could better capture the decision process of the non-linear operation.

Even though LR is not as used in the reality of predicting stock market performance and market trends, it has shown that it can be compared to other more proficient algorithms in forecasting outperforming stocks. (Gong and Sun, 2009) proposed an innovative model by suggesting the use of Logistic Regression to predict the stock market direction of the following month, using market indexes and information of the ongoing month. This approach showed promising results reaching an accuracy value of 83.3 %. (Upadhyay et al., 2012) combined Multinomial Logistic Regression (MLR) model with seven different financial ratios to predict stock performance and classify it according to stock return. Their results showed an accuracy rate of 56.8 % when dividing stocks into three categories. (Ali et al., 2018) developed a Logistic Regression model to forecast stock performance and classified them as

good or bad according to a cut-off value defined. Combining with financial ratios they obtained very encouraging results with an 88.4 % overall accuracy.

Furthermore, there are cases where LR has been computed with no connection to the financial sector at all. Examples of this are (Lee, 2004) and (Lee et al., 2007), who used Logistic Regression with the purpose of landslide susceptibility mapping in South Korea. This model showed different results depending on the areas of the study being east Korea or west Korea, and the classes of accumulated area ratio.

Although not all studies presented above are precisely connected with the purpose and the variables of this work, every single one provides an idea of the capacities of the Logistic Regression model and the diverse applications of the method. Most importantly, they serve as foundations for the work due to the promising results in the areas described.

2.8.3 Support Vector Machine

SVM is a very known machine learning algorithm mostly used in classification and regression problems. Due to advancements of the models over the years, SVM has been used for all types of problems, linear or non-linear.

The foundation of the Support Vector Machine technique was advanced by Vladimir Vapnik (Vapnik and Lerner, 1963) to serve as a linear classifier. (Boser et al., 1992) developed the initial model to create a non-linear classifier. They suggested an approach to maximize the margin of the hyperplane by applying the ”kernel trick”. This approach was implemented using different classification functions and showed good generalization prowess when compared to other learning algorithms. When dealing with non-linearly separable data, (Cortes and Vapnik, 1995) proposed the soft margin implementation, allowing a few cases of margin breach.

More recently, SVM has been applied to various problems, all with different implements of the algorithm. The model uses heuristics to solve Quadratic Programming (QP) problems, the base of the optimization. (Wang and Hu, 2005) changed the base of the problem by creating a Least Squares version of SVM (LS-SVM), that uses equality constraints and a sum squared error cost function instead of the QP problem. This version was used to estimate non-linear functions and, compared to a normal SVM, achieved better results in large-scale regression problems.

(Han and Chen, 2007) combined an SVM model, using a gaussian radial basis function as a kernel function, with ratio analysis, by selecting several different financial ratios, with the purpose of predicting stocks based on their level of profitability. They concluded that earnings per share and book value per share used as features obtained the best results. On the topic of predicting future financially distressing companies, (Xie et al., 2011) created an SVM and an MDA model to be used in the Chinese stock market. Combining the models with financial ratios as well as governance indicators and market variables, the SVM model outperformed the MDA model, achieving more than 80 % accuracy in predicting

companies three years prior to the financial incidents.

Non-related to the financial markets, (Jottrand, 2005) developed an SVM model to recognize facial expressions and detect camouflages in a vegetation environment. On the topic of facial recognition, the model proved to be very flexible, providing good results even with a low-level normalization algorithm. Regarding the camouflage detection, the model was able to detect most of the nets having a hard time on the small ones which are more precise. The only problem shown by the SVM technique was concerned with the false alarms depending on the vegetation environment selected. (Noori et al., 2011) investigated the performance of SVM and ANN on predicting the monthly streamflow in the Sofichay River in Iran. They combined the two models with a PCA and a Gamma Test (GT) to reduce the dimensionality of the inputs. They concluded that the PCA and GT techniques improved the results obtained by a good margin when compared to a simple SVM model, but also that the SVM model proved to be more accurate, than the ANN model, at predicting the monthly streamflow.

2.9 Conclusions

This chapter contains explanations of notions important to this work. First, an introduction to the stock market is made, with explanations about movements of the markets and approaches to invest. Second, it is illustrated one of the most important investment techniques utilized by investors so far, which is fundamental analysis, where all the components that are related to this approach are described e.g., the financial analysis executed based on the financial statements, the details that analysts have to be aware when performing industry analysis, and the overall idea that financial ratios give of the relative strength of companies. All these components are valuable pieces, for analysts and investors, to achieve a better perspective of the evolution of the market and profitability of their investments.

After, some basic machine learning concepts are introduced, such as overfitting and cross-validation, which are notions that have to be considered when using a machine learning algorithm. Then, the algorithms are explained. The topics of Logistic Regression, Support Vector Machines, and Principal Component Analysis are analyzed, with brief explanations to how to implement them and notions about necessary parameters. Finally, performance measures are explained, which are used to evaluate the models, as well as the literature review of every main topic in this work, presenting different researches where these topics have been applied.

Although the combination of fundamental analysis and Logistic Regression is not very common in the world of predicting stock market evolution, each one has its share of researches based on. FA has been of great use for investors in this kind of subject showing very promising results. (Yu et al., 2009) discovered that three out of the four key determinants of stock index movement were fundamental variables. LR has been mostly used in the subject of bankruptcy where the goal is more suited to a classifier algorithm, with very few exceptions of papers where LR was applied to stock performance prediction showing an above 80 % accuracy in some cases.

Table 2.2: Summary of existing approaches with Fundamental Analysis.

Reference	Year	Data input	Algorithm	Period	Results
Altman	1968	Financial ratios derived from company's financial statements	MDA	1946-1965	95 % accuracy on bankruptcy prediction
Mubin et al.	2014	Financial ratios : Profit Margin, Assets Turnover and Equity Multiplier	ANOVA and Regression Analysis	2004-2012	Assets Turnover showed high volatility among different industries unlike the other ratios
Fama and French	1992, 2015, 2016, 2017	Financial ratios: firm size, book-to-market equity, excess return on the market, profitability and investment factors	cross-sectional regression	1962-1989	size and book-to-market equity capture the average stock returns associated to size, E/P, book-to-market and leverage
Olson and Mossman	2003	Financial ratios	NN, OLS and LR	1973–1993	NN outperforms both OLS and LR in stock returns prediction
Albanis and Batchelor	2007	Financial ratios	LDA, LVQ, PNN, OC1 and RI	1993–1997	Unanimous voting principle outperforms the majority voting one
Silva et al.	2015	Financial ratios combined with technical indicators	MOEA	2010-2014	Increase of fundamental ratios can increase returns and precision
Tsai et al.	2011	Financial, economical and technical ratios	NN, DT and LR	2002-2006	classifier ensembles outperform single classifiers on return and accuracy

Table 2.3: Summary of existing approaches using Logistic Regression.

Reference	Year	Data input	Algorithm	Period	Results
ohlson	1980	Information from financial statements	LR	1970-1976	Dependance on information available
Zavgren	1985	Financial ratios	LR	1972-1978	Model showed good results in detecting signs of ailing companies
Maher and Sen	1997	Financial ratios	NN and LR	1962-1978	Neural networks outperforms Logistic Regression in bond rating prediction
Min and Jeong	2009	Financial ratios	NN, DT, LR	2001-2004	GA showed 79 % accuracy against LR's 71 %
Chen	2011	Financial and non-financial ratios	DT and LR	2000-2007	DT showes better results for near collapsing firms while LR ouperformed in a distant future
ogut et al	2009	index's average daily return, average daily change in trading volume and average daily volatility	ANN, SVM, LR and DA	1995-2004	ANN and SVM are more suited to detect stock manipulation
Lee, Lee et al.	2004, 2007	Geographic information system (GIS)	Likelihood ratio, LR and ANN	———	Each algorithm showed better results in different geographic areas
Gong and Sun	2009	Feature index variables	LR and ANN	2005-2007	Both algorithms showed accuracy levels of 83%
Upadhyay et al.	2012	Financial ratios	MLR	2005-2008	the financial ratios picked with LR model reached an accuracy of 56.8 %
Ali et al.	2018	Financial and accounting ratios: ROE, CR, debt to equity, EPS, sales growth and book to price ratio	LR	2011-2015	The model showed an 88.4 % overall accuracy in classifying good and bad companies

Table 2.4: Summary of existing approaches using Support Vector Machines.

Reference	Year	Data input	Algorithm	Period	Results
Boser et al	1992	hadnwritten digits and mail pieces	SVM	———-	”Kernel trick” improved the generalization ability and should be considered as a cost function optimization option
Wang and Hu	2005	non-linear functions	LS-SVM and SVM	———-	LS-SVM model has the upper hand in large scale regression problems than a regular SVM model
Han and Chen	2007	Financial ratios: Earnings per share, book value per share and net profit growth rate	SVM	———-	SVM with earnings per share and book value per share showed the best results
Xie	2011	Financial ratios with governance indicators and market variables	SVM and MDA	2002-2007	SVM showed an 80 % accuracy in predicting financially distressing companies
Jottrand	2005	pixelated facial and vegetation pictures	SVM	———-	Great flexibility shown by the SVM model in the tests performed providing high accuracy rates
Noori et al.	2011	weather variables	SVM and ANN	1983-2004	feature selection techniques improve results

Chapter 3

Implementation

The goal of this project is to test the competitiveness of a system based on Logistic Regression with the use of Fundamental Analysis. A SVM model was also designed for comparison purposes and to properly evaluate the performance. This chapter will explain the reasoning behind the implementation, as well as fully describe all the structure developed.

In this chapter, a description of the developed implementation is presented based on the information already outlined in the previous chapter. Section 3.1 provides an idea of the overall system, describing the goal behind each layer. Section 3.2 describes the necessary information for this work and the methods to retrieve it. Section 3.3 gives an overview of the financial ratios and labels that will be used in the system, how to calculate them and what implications to take from them. Section 3.4 explains the training procedure of the system, describing the parameters, methods to tune them and how to evaluate the models. Section 3.5 outlines the testing procedure, where the predictions are performed by the models. Finally, Section 3.6 provides real insight into the way investments are made, describing every strategy performed and the ways to evaluate results.

3.1 System Architecture

This work presents a combination of a Logistic Regression model with Fundamental Analysis to predict which companies will have major increase in value with the goal of maximizing profits. Basically, the model takes the raw data from each firm, computes selected ratios to evaluate performance which will then result in the best companies to invest according to the model. Based on the companies returned, a simulation of the investments will be made to evaluate the performance of the system. It is also important to mention that a Support Vector Machine model will be constructed as well in order to compare performances. As illustrated in Figure 3.1, the system's architecture is divided in five layers. Each layer has its role in the system and is independent from one another, in the sense where it provides flexibility to the system by being able to be run separately, and allowing the addition of new modules if necessary. Each layer's responsibility can be seen below:

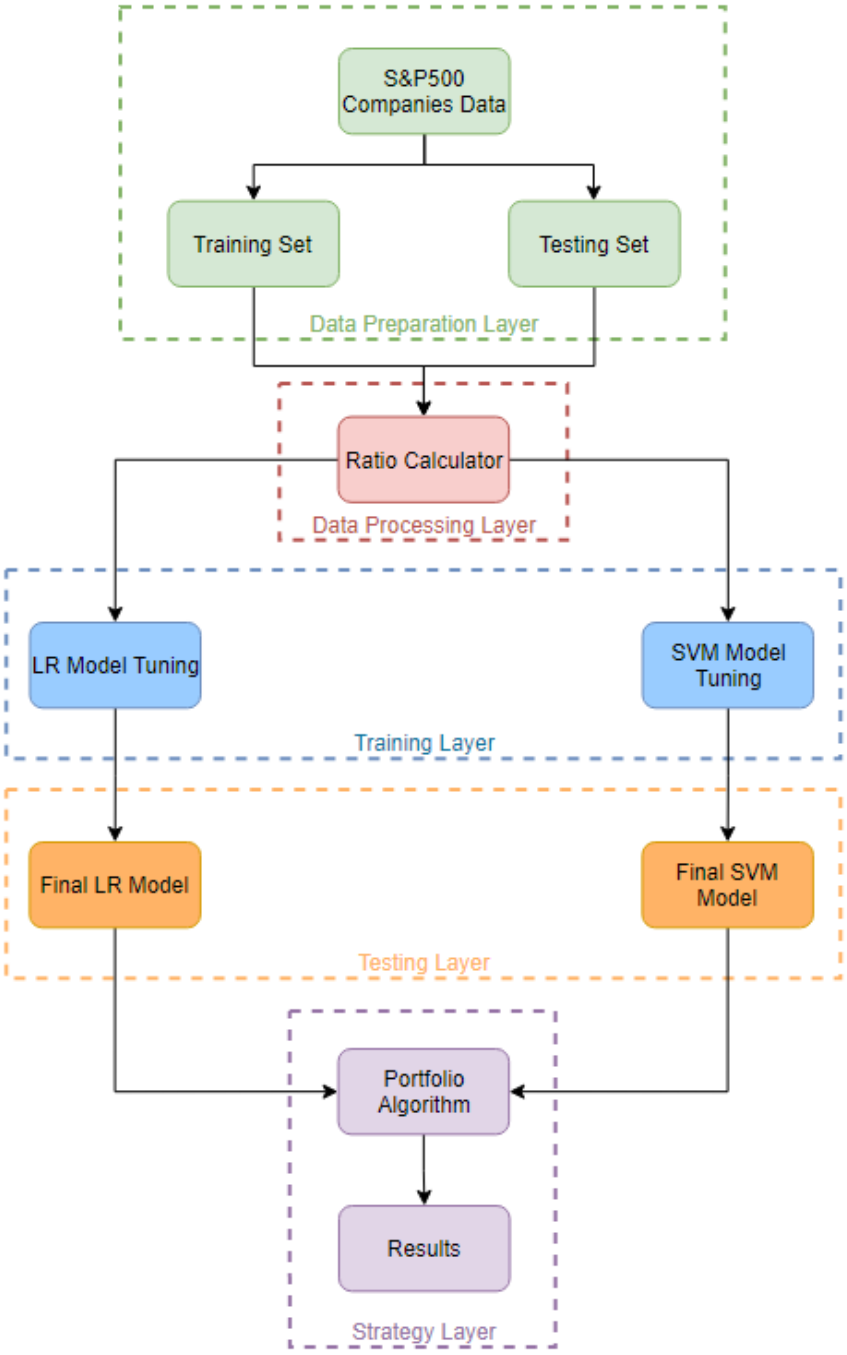


Figure 3.1: Flowchart of the overall system.

The **Data Preparation Layer** is in charge of obtaining the data necessary for model, filter it and prepare it for processing. This information, as stated before, comes from the financial statements where only the required indicator values are kept. Essentially, it retrieves all the necessary fields of the desired companies, guarantees the completeness of the information and returns the training and testing sets of raw data as output.

The **Data Processing Layer** uses the datasets of raw data as inputs and computes the financial ratios and the labels with the information. In the end, it will modify the datasets in order to have the ratios desired so that the models created can be trained and tested.

The **Training Layer** was created with the purpose of tuning the models described. Since both Logistic Regression and Support Vector Machine have parameters to be optimized, an extensive search has to be made in order to guarantee an appropriate model for each circumstance. Ultimately, the models use the training set to evaluate a number of parameters and return the set of parameters that obtains the best result of the metrics chosen.

The **Testing Layer** takes the set of parameters returned from the training layer and applies the resulting model to the testing set. With this, it obtains the confidence level on each company to grow in the near future. This list will then be sent to the strategy layer.

The **Strategy Layer** receives the list of companies and evaluates the model based on the results obtained and profits made by the companies selected. Different investment strategies are tried and compared to have a better insight into the behaviour of each choice.

Each layer mentioned will be described in greater detail in the following sections.

3.2 Data Preparation Layer

The data preparation layer gathers all data that will be used by the data processing, training and testing layers. It is divided in three modules: compustat database, macrotrends and yahoo finance and individual years. The first one represents the data from 2005 until 2018 from the companies, which is already stored. To this, the information from 2018 onward will be added finishing the training set. The final module represents the data for each year of the companies in the index, making the testing sets. The datasets were divided to facilitate the work and to be able to only evaluate on the companies present in each year. An overall architecture of this layer is presented in Figure 3.2.

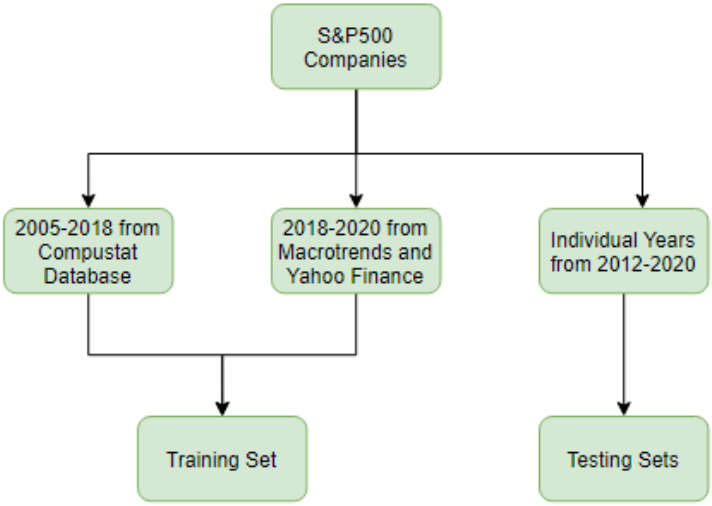


Figure 3.2: Architecture of the Data Preparation Layer.

This financial information is contained in the financial statements issued by the companies, which are the income statement, balance sheet and cash flow statement. Macrotrends and Yahoo Finance are two databases that allow users to download public information about a company's market and its financial statements, whether they are quarterly or annually. From the modules displayed in Figure 3.2,

the data from the Compustat database was already present in an excel file ready for processing, but the modules from 2018-2020 and from 2012-2020 were still missing. With this in mind, two programs were developed in python, represented in Figure 3.3, in order to retrieve this information and store it in pickle files.

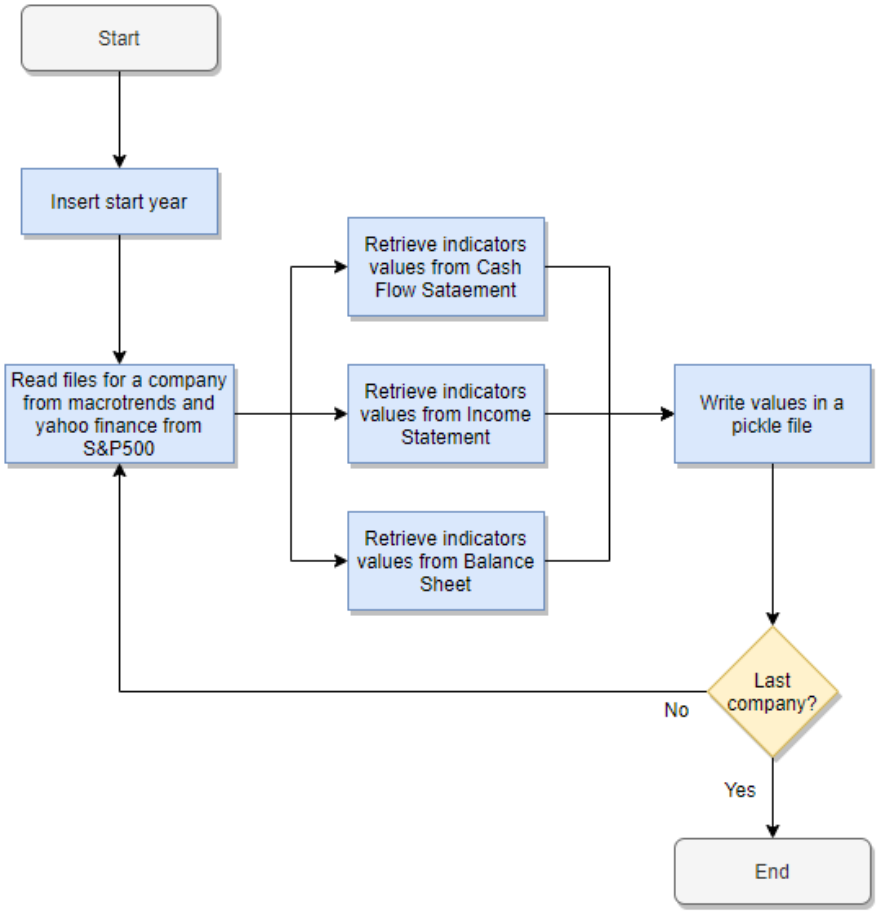


Figure 3.3: Flowchart of the program developed to retrieve financial data.

The companies selected for this work are the ones included in the Standard and Poor's 500 (S&P 500), since they represent a good sample of different industries with leading companies. The companies were filtered in regards to the amount of information available, using a python module named *pandas* known for its use in manipulating data within python, being only kept the ones with no missing data over the years. The information collected represents a period between 2009 and 2021 illustrating different financial situations, from crisis to economic booms, and it is a big enough sample for diverse training and testing sets.

3.2.1 Financial Data

The raw data retrieved will then be turned into ratios in the data processing layer. Due to this reason, the module acquires only the data specified by the user in order to later obtain the desired ratios. With this being said, the following fields were selected:

- **Revenue -** Total amount of income produced by the sale of goods or services associated to the company's primary business.
- **Cost of Goods Sold -** Total cost of producing the goods that the company sold.
- **Gross Profit -** Profit made by a company after deducting the cost of producing and selling its goods.
- **Income Tax -** Estimate of how much a company pays in taxes in an accounting period.
- **Net Income -** Total amount of money a company is making after deducting all expenses and costs from revenue.
- **EBIT -** Profit made by a company before applying taxes and interest.
- **Earnings per Share -** Portion of a company's profit given to each outstanding share of common stock.
- **Total Assets -** All items of value owned by a company.
- **Total Current Assets -** All assets that have an expectation of being converted to cash in a one year period.
- **Inventory -** All items controlled by a company to sell and make a profit.
- **Pre-Paid Expenses -** Assets that result from advanced payments made by a company for goods to be received in the future,
- **Total Liabilities -** Total amount of debts and obligations that a company owes to outside parties.
- **Total Current Liabilities -** All liabilities that are due within one year or operating cycle.
- **Shareholder Equity -** How much owners of a company have put into the business.
- **Book Value per Share -** Minimum value of a company's equity.

This data is stored in a dataframe, with the Apple (APPL) data represented in Figure 3.4 as an example, for an easy-to-use data manipulation. This information will then enter the ratio calculator module for processing.

datadate	Revenue	Of Goods	Gross Profit	Income Taxes	Net Income	EBIT	Earnings Per	Current A	Total Assets	Inventory	Paid Expen	Current Liab	Total Liabilitie	are Holder Equ	ook Value Per Shar	tic	divy
2018-03-31 00:00:00	61137	37715	23422	2346	13822	15894	0.6825	130053	367502	7662	0	89320	240624	126878	6.4167	AAPL	0.1762
2018-06-30 00:00:00	53265	32844	20421	1765	11519	12612	0.585	115761	349197	5936	0	88548	234248	114949	5.9339	AAPL	0.1762
2018-09-30 00:00:00	62900	38816	24084	2296	14125	16118	0.74	131339	365725	3956	0	115929	258578	107147	5.6334	AAPL	0.1762
2018-12-31 00:00:00	84310	52279	32031	3941	19965	23346	1.045	140828	373719	4988	0	108283	255827	117892	6.2313	AAPL	0.1762
2019-03-31 00:00:00	58015	36194	21821	2232	11561	13415	0.615	123346	341998	4884	0	93772	236138	105860	5.7442	AAPL	0.19
2019-06-30 00:00:00	53809	33582	20227	1867	10044	11544	0.545	134973	322239	3355	0	89704	225783	96456	5.3215	AAPL	0.19
2019-09-30 00:00:00	64040	39727	24313	2441	13686	15625	0.765	162819	338516	4106	0	105718	248028	90488	5.0913	AAPL	0.19
2019-12-31 00:00:00	91819	56602	35217	3682	22236	25569	1.25	163231	340618	4097	0	102161	251087	89531	5.1044	AAPL	0.19
2020-03-31 00:00:00	58313	35943	22370	1886	11249	12853	0.6375	143753	320400	3334	0	96094	241975	78425	4.5343	AAPL	0.3556
2020-06-30 00:00:00	59685	37005	22680	1884	11253	13091	0.645	140065	317344	3978	0	95318	245062	72282	4.2182	AAPL	0.3556
2020-09-30 00:00:00	64698	40009	24689	2228	12673	14775	0.7475	143713	323888	4061	0	105392	258549	65339	3.8487	AAPL	0.3556
2020-12-31 00:00:00	111439	67111	44328	4824	28755	33534	1.68	154106	354054	4973	0	132507	287830	66224	3.9365	AAPL	0.3556

Figure 3.4: Data format of APPL stock.

3.3 Data Processing Layer

The data processing layer has the task of preparing the features to be used by the LR and SVM models in the following layers. The data described in 3.2.1 enters the ratio calculator module, where a series of arithmetic operations occur, computing the ratios desired. Finally, a verification of the data is done. This verification is due to the fact that certain ratios depend on certain factors being true. For example, the inventory turnover ratio depends on the company having inventory which is not always the case. So neutral values have to be put in certain places to guarantee the correctness of the data and make sure that the models can work without misleading values.

These features will be used for prediction by the models in the training and testing layers. The ratios are more deeply described in section 3.3.1 while in section 3.3.2, the labels of the data will be explained.

3.3.1 Financial Ratios

Using the information from the financial statements defined, financial ratios are calculated with the purpose of analyzing a company regarding liquidity, leverage, efficiency, profitability and market value. These ratios are then used to compare companies included in the same sector and choose the right companies to invest.

Liquidity ratios are financial measures that represent a company's ability to pay off its current debts without requiring external money. Table 3.1 contains the liquidity ratios used in this research.

Table 3.1: Liquidity Ratios.

Ratios	**Formula**	**Implications**
Quick Ratio	$\frac{\text{CE + MS + AR}}{\text{Current Liabilities}}$	QR = 1 means a company can cover its liabilities. A higher value tends to indicate a healthy company
Current Ratio	$\frac{\text{Current Assets}}{\text{Current Liabilities}}$	CR $<$ 1 means a company can't cover its short-term debts. A higher value tends to indicate a healthy company

Leverage ratios assess the capacity of a company to satisfy its financial obligations. It is a way to track where the capital is coming from. In table 3.2, the leverage ratios picked are presented.

Table 3.2: Leverage Ratios.

Ratios	**Formula**	**Implications**
Debt-to-Equity	$\frac{\text{Total Liabilities}}{\text{Shareholders' Equity}}$	A low value tends to indicate less risk for shareholders. Optimal amount of debt depends on the industry
Interest Coverage	$\frac{\text{EBIT}}{\text{Interest Expense}}$	A value lower than 1.5 may indicate problems in meeting interest expenses

Efficiency ratios measure a company's performance by analyzing its ability to use its assets to make

money. Table 3.3 presents the ratios used.

Table 3.3: Efficiency Ratios.

Ratios	Formula	Implications
Asset Turnover	$\frac{\text{Total Revenue}}{\text{Average Assets}}$	The greater the value, the more effectively a company is generating revenue from its assets
Inventory Turnover	$\frac{\text{Cost of Goods Sold (COGS)}}{\text{Average Inventory}}$	A low value tends to indicate weak sales. A higher value can either imply strong sales or insufficient inventory

Profitability ratios represent the effectiveness of a company on generating profit for its shareholders. Table 3.4 contains the profitability ratios used in this work.

Table 3.4: Profitability Ratios.

Ratios	Formula	Implications
Return on Equity	$\frac{\text{Net Income}}{\text{Shareholders' Equity}}$	An acceptable value depends on the industry average. Anything under that value is considered poor
Return on Assets	$\frac{\text{Net Income}}{\text{Total Assets}}$	Like ROE, it is important to compare with similar companies. A higher value tends to indicate a more earning company with less investment
Gross Profit Margin	$\frac{\text{Total Revenue - COGS}}{\text{Total Revenue}}$	A good value will depend from industry to industry. A high value indicates that there is more money to reinvest and profit from
Operating Profit Margin	$\frac{\text{EBIT}}{\text{Total Revenue}}$	Analyzing evolution over a period is essential. A high value shows that a company is making its profit from core operations

Market value ratios measure the share price of a company's stock with the purpose of evaluating if it is over-valued or under-valued. In table 3.5, the market value ratios chosen are presented.

Table 3.5: Market Value Ratios.

Ratios	Formula	Implications
Price-to-Earnings	$\frac{\text{Market Value per Share}}{\text{Earnings per Share}}$	A low P/E can be an indication of an undervalued stock while a higher value can mean investors are expecting a high earnings growth
Price/Earnings-to-Growth	$\frac{\text{Price-to-Earnings}}{\text{Earnings per Share Growth Rate}}$	A value of 1 means a company is fairly valued which means a PEG < 1 might be an indication of an undervalued stock
Price-to-Book	$\frac{\text{Market Value per Share}}{\text{Book Value per Share}}$	A P/B value under 1 generally means that the stock is undervalued and the company might be a solid investment
Earnings per Share	$\frac{\text{Net Income - Preferred Dividends}}{\text{Average Outstanding Shares}}$	The greater the EPS value, the more profitable a company is considered to be
Dividend Yield	$\frac{\text{Dividend per Share}}{\text{Market Price per Share}}$	A high value of DY might mean that the company is paying its shareholders a good amount.
Dividend Pay-out	$\frac{\text{Dividend per Share}}{\text{Earnings per Share}}$	A high percentage indicates a company that is paying their entire net income as dividends. A low value means it is retaining the money for reinvestment

It is important to note that all the implications presented in the table above are highly dependant on the industry of each company, management strategy and other variables. This way, it is almost impossible to draw any conclusions by looking at very few ratios. That is why, in this research, over 15 ratios are calculated with the intention of letting the algorithm pick the most important ones, and the ones that can capture best the financial situation of companies.

3.3.2 Labels

Since both Logistic Regression and Support Vector Machine are supervised learning algorithms, a label is needed for the training procedure of the system. This label represents the very thing the algorithm is trying to predict correctly. However, in the training environment, this label is necessary so that the model can distinguish between each group of data.

Considering the purpose of this work is to find companies with a steady long-term growth, the label has to represent exactly that. However, a label based just on a growth number is not the perfect strategy since it does not take into consideration market factors nor gives real insight into the potential of the investments. Due to all this reasons, the label chosen represents if the company had outgrown the S&P 500 index value from financial period to financial period with a value of 1 and a 0 if the premise is not true. This guarantees, if the predictions are correct, that the model has good potential providing out-performing results and becoming a noticeable research for the community.

3.4 Training Layer

The training layer is tasked with training each model in order to find the one that guarantees the best results in the testing phase. A brief look of the layer can be seen in Figure 3.5. First, the sector and year are selected. This is done since each industry has very different implications on financial data. So, for this research, each sector will have a different model by year in order to be as reliable as possible. Second, the data is scaled. Even though the information has been filtered and outliers have been erased, the data is scaled to further guarantee quality of the data and to improve results by not allowing particular features to impact the predictions too much. Third, the grid search algorithm is implemented to perform an extensive search through the parameters of each model. For the Logistic Regression model, one search was performed for each solver, totalling five due to the high range of parameters to tune. For the Support Vector Machine model, one search per kernel is also done although the range of parameters is different. Lastly, a model for each sector and year is selected and sent to the testing layer in order to obtain final results.

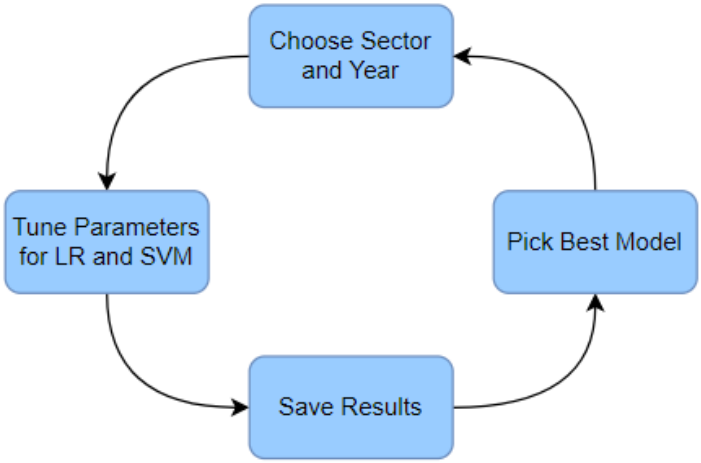


Figure 3.5: Architecture of the Training Layer.

3.4.1 Hyperparameters

The first parameter to tune in the LR model is the solver, as described in section 2.4.1. Each solver solves the optimization problem differently and so, for different kinds of data, different results will be obtained. Another parameter is regularization. Different types of penalties are supported by each solver. These penalties are used to decrease the error by fitting a function on the training data. The parameter C is the inverse of the regularization strength, the higher it is, the bigger the weight given to the data. This leads to greater penalization for incorrectly classified samples, which, if not properly adjusted, can result in overfitting. Last parameter tuned is class weight. Class weight penalizes mistakes of each class with its corresponding weight. A higher value means it is giving more emphasis to a specific class than the other.

In regards to the SVM model, and similar to the LR one, the first parameter tuned is the kernel,

mentioned in section 2.5.2. The linear kernel is simpler than the rest, being faster to train but being limited when the data is not linearly separable. The polynomial, radial basis function and sigmoid kernels can take non linear problem approaches with the polynomial being the most versatile. Once again similar to the LR model, the parameters C and class weight are also tuned for the SVM model, having the exact same functions as the ones described before, but having a different range of values. Last but not least, the Gamma parameter. This parameter, which doesn't exist for the linear kernel, influences the curvature of the decision boundary, with a higher value resulting in a bigger curve. As the other parameters, a value too big may result in overfit although it always depends on the data.

3.4.2 Grid Search

With the goal of finding the best parameters for the data, a grid search algorithm is applied. This algorithm will train and test both the LR and SVM models several times with different values of all the parameters described in section 3.4.1. The solutions are then evaluated and one set of hyperparameters is chosen in the end. Since this algorithm is performed for each sector in every year, a high amount of models are saved and then used accordingly.

Cross-Validation, described in section 2.3.3, is also used to make sure the solutions found are not overfitted. A 5-Fold validation is applied to almost all sectors' data so that each folders can still have a significant size of training data, and the tests are a significant number as well avoiding biased solutions. A 3-Fold is used for the sectors which have less data, in this case only the Real Estate sector, so that the folders won't be reduced too much and the tests can still have meaningful insight.

3.4.3 Fitness Function

The fitness function used in this work was the ROC-AUC. As explained in section 2.7, ROC-AUC uses the probability outputs of the model and presents the tradeoff between the true positive rate and the false positive of predictions at various classification thresholds as a curve. At each threshold, the system is evaluated, where both a higher true positive rate and a lower positive rate mean that the system can properly distinguish between classes, which will result in a larger AUC.

In the context of this work, ROC-AUC was picked since it examines how well the model ranks predictions, meaning that a high fitness value is related to a bigger probability of ranking a positive observation with a higher value than a negative one.

3.5 Testing Layer

The testing layer is probably the simplest part of the entire system. It takes the models that obtained the best solutions in the training layer and uses them on the not yet seen testing data. Here, the percentages of the predictions for each company are calculated and stored in a dataframe. These percentages represent the confidence level of the model in a company outgrowing the S&P 500 index for that financial period. Since the purpose of this work is to find companies with longevity potential, an average of the percentages is done, so that one value represents the potential of a company over a

year, according to each model. Finally, the list of companies with their respective percentages is ordered so that the highest values can be selected. The selected companies are then sent to the portfolio layer where the investment will be performed and evaluations will be done.

3.6 Portfolio Layer

The portfolio layer is responsible for finding a way to make money with the investments on the companies provided by the LR and SVM models. Two different lists are provided by the testing layer, one regarding the top 20 companies according to the model, and the other regarding the companies with a confidence level above 60 %, making use of the percentage results. Then, the daily prices are retrieved and the daily returns are computed as well as other variables like 200-day moving average, an increasing price streak vector and a decreasing price streak vector. These variables will be used by the module in the investment strategies applied. Having all the data ready, five different investment strategies are implemented, which will be further explained below, regarding the logic behind buy and close orders in the stock market. This will provide good insight of each company's potential and the best way to interact with the stock market. For this, an investment vector is created, where, on the upcoming year to the test set, the module indicates in the vector, the days when the model is in or out of the market. Ultimately, an evaluation of the results is performed in order to draw conclusions of the whole system, the models designed and the investment strategies implemented.

3.6.1 Strategy 1

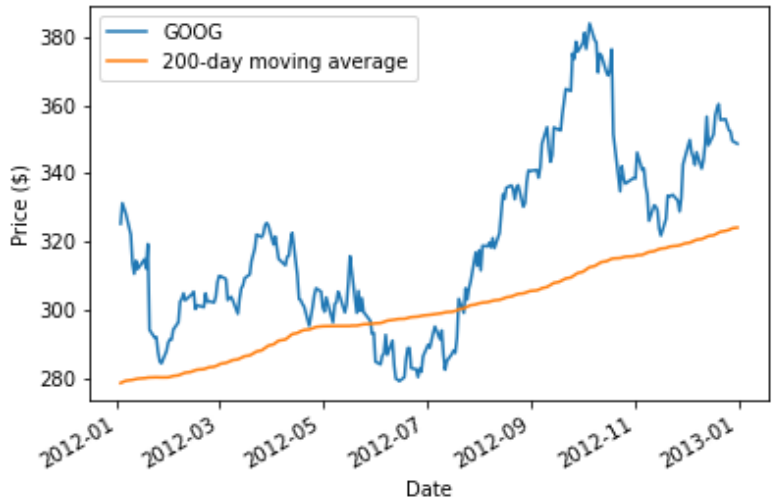
The first implemented strategy, which from now on will be referred to as MA200, takes into consideration the 200-day moving average. This moving average represents the average price over the 200 prior days and gives an idea of the price trend, whether it is up or down. The module uses this average to determine when to buy or sell, checking if the daily price is higher than the average to stay in and leave if it decreases the average price. An example of the functioning of the strategy can be seen in Figure 3.6.

Using the Google company's share (GOOG) as an example, it is possible to observe the goal of the implementation. After the investment made, in Figure 3.6(a), when the price drops and crosses the moving average, the system closes the position only opening a new one when there is a new upwards cross. The goal is to reduce the losses in cases of decline and preserve profits in cases of an increase, demonstrated by Figure 3.6(b).

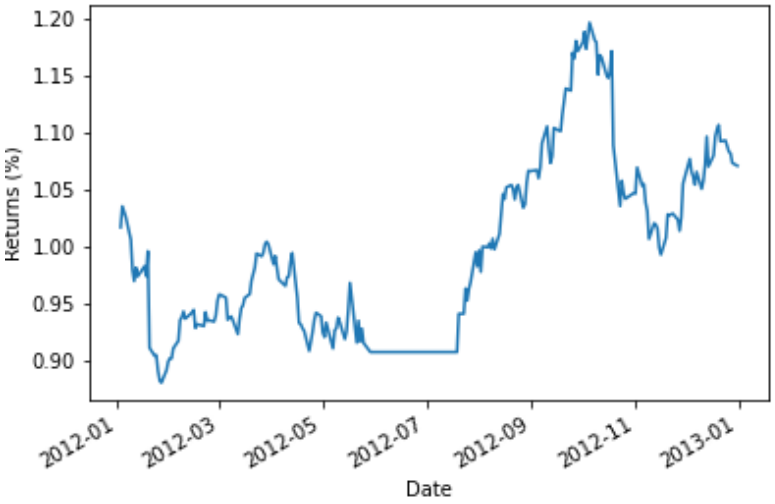
This approach can have very different outcomes on the amount of trades performed depending on the price trends. An upward trend can lead to only one long position being opened while a sideways type of market will lead to a high amount of buy and sell orders from the algorithm.

3.6.2 Strategy 2

The second scenario, which from now on will be referred to as Stoploss, only takes into consideration the variation of the prices, by computing a stop-loss variable. This stop-loss is responsible for detecting if the



(a) Stock Price and Moving Average



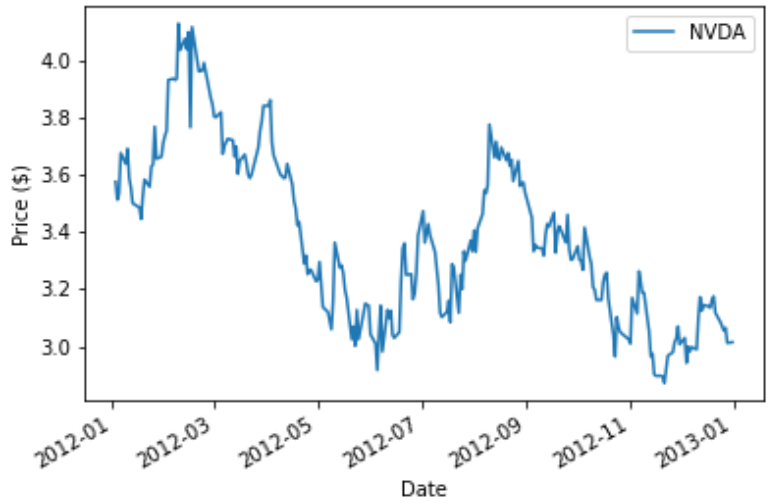
(b) Cumulative Returns

Figure 3.6: MA200 summary.

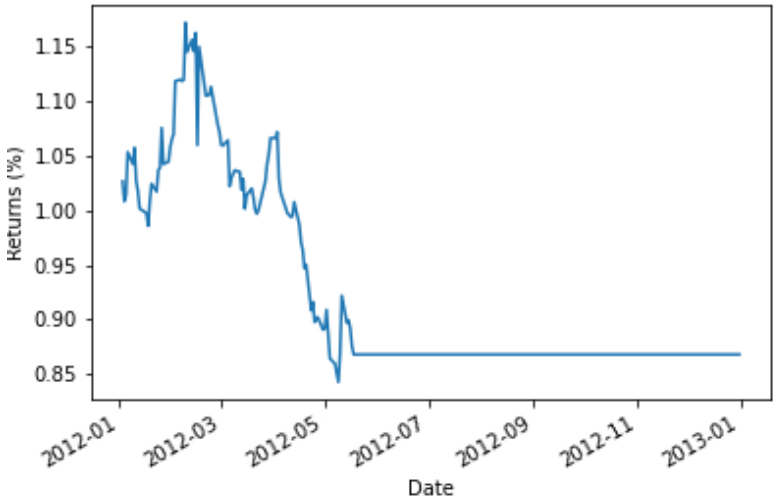
price has decreased over a threshold, in this case a 15 % decrease, closing the position if the condition is true. This technique is used to prevent the losses in graphs that do not show signs of recovery while keeping sideways markets and upward trends intact. An example of the functioning of the strategy can be seen in Figure 3.7.

Using the Nvidia company's share (NVDA) as an example, the system evaluates at all times if the price has dropped over the threshold defined at the moment of investment. In this particular case, the loss is reduced due to the close order performed preventing an even bigger deficit, as shown in Figure 3.7(b).

This strategy leads to a very low amount of trades, in fact, if the condition is not met, only a buy order is made with a closing order at the end of the testing period. With the activation of the condition, the closing order is performed earlier in the year.



(a) Stock Price

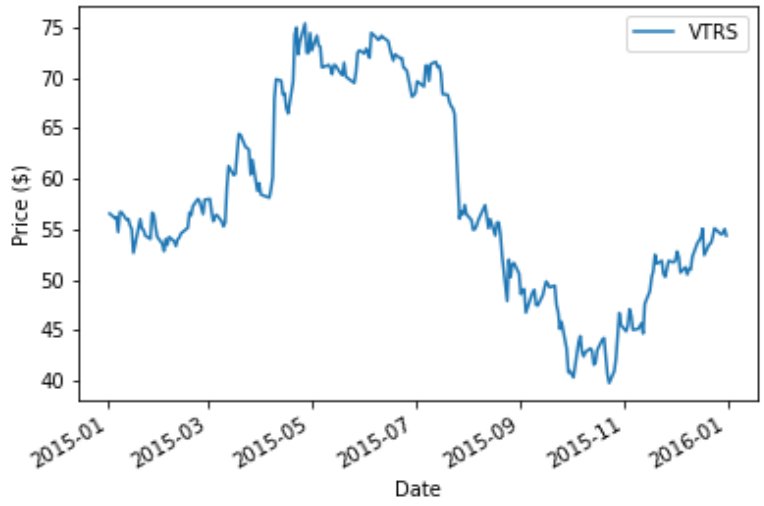


(b) Cumulative Returns

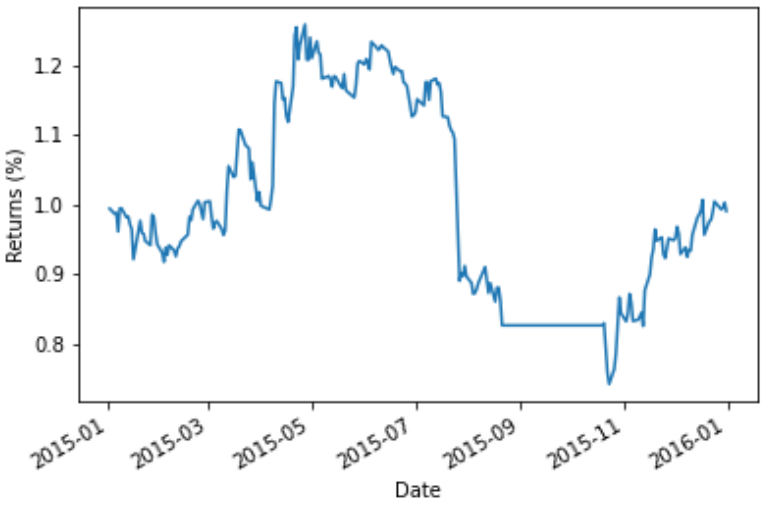
Figure 3.7: Stoploss summary.

3.6.3 Strategy 3

The third strategy, which from now on will be referred to as StoplossReentry, is very similar to the second one, with a stop-loss threshold implementation, but with an addition, an option to re enter the market after the stop-loss trigger. The threshold functionality is exactly the same as the one explained before while the market re entrance is evaluated according to the latest prices. Basically, the system evaluates if the graph is recovering with an upward trend and, if that is the case, it opens a new position. This functionality is performed by using a 15-day price differential and a 3-day streak, checking if the price difference crosses above 5 % while maintaining a positive streak. This is computed in order to find moments where the second strategy misses, where a stock goes down but ends up recovering and ending on a high. An example of the functioning of the strategy can be seen in Figure 3.8.



(a) Stock Price



(b) Cumulative Returns

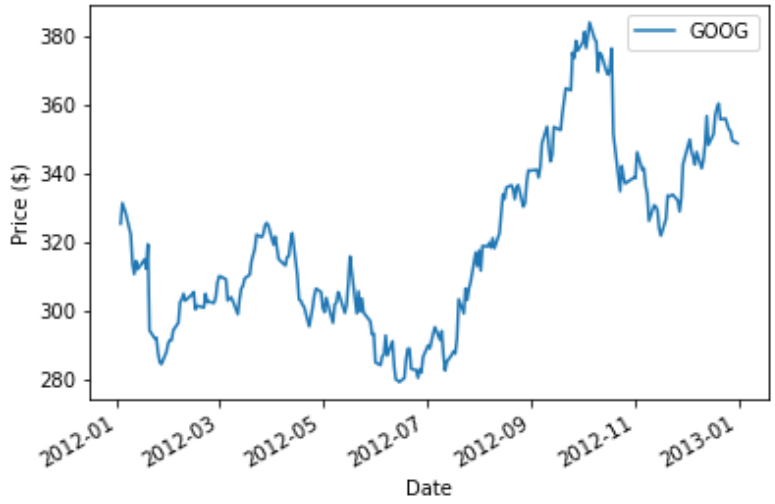
Figure 3.8: StoplossReentry summary.

Using the Viatris company's share (VTRS) as an example, the system evaluates at all times if the price has dropped over the threshold defined at the moment of investment, as does strategy two. However, after the threshold is triggered, the algorithm looks for a way back in, in order to try and recover the losses obtained. Figure 3.8(b) shows how the algorithm managed to recoup some of the losses and turn into profit.

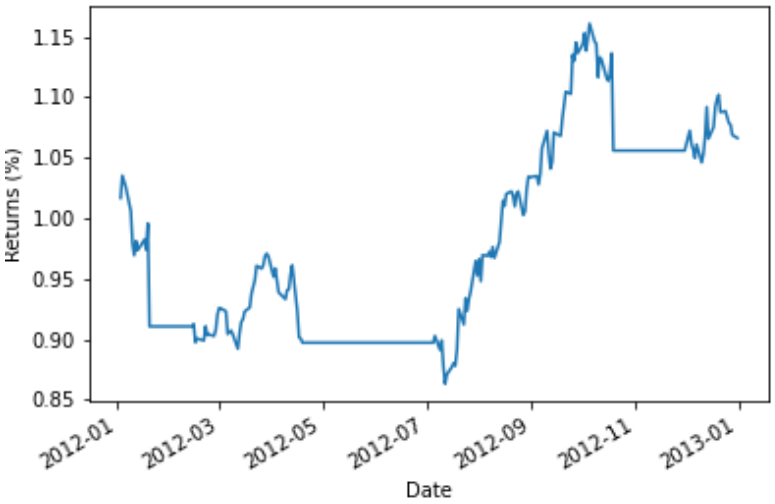
Similarly to strategy number two, the number of trades performed is very low, with a maximum of two buy orders and two sell orders counting the automatic close order at the end of the testing year.

3.6.4 Strategy 4

This strategy, which from now on will be referred to as Momentum, is the more complex of the five. It is a momentum based strategy that takes into consideration the 15-day price differential, the upward and downward 3-day price streak and the stop-loss threshold. This strategy works similar to the re buy functionality implemented in strategy three. In order to close positions, the algorithm checks for the stop-loss threshold, for a steady decline of price value over a number of days or a sudden considerate drop in stock price. In order to open positions, the algorithm checks for upward trends following closed positions with the goal of cutting the losses and recovering quickly. An example of the functioning of the strategy can be seen in Figure 3.9.



(a) Stock Price



(b) Cumulative Returns

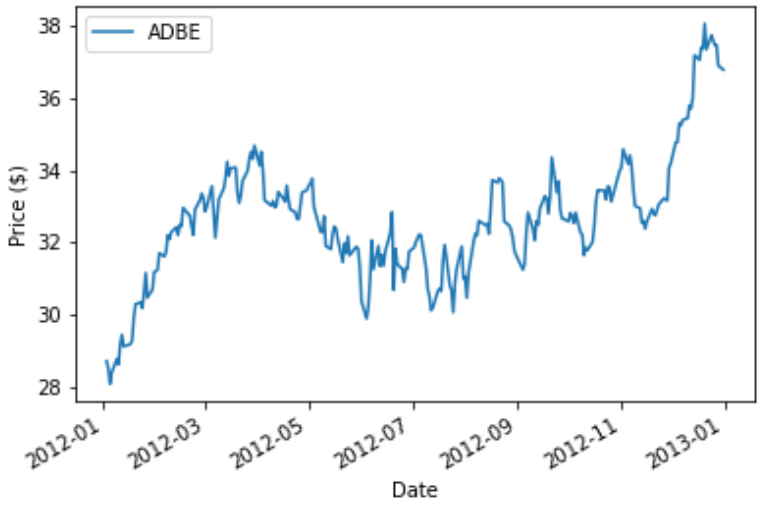
Figure 3.9: Momentum summary.

Going back to the data from Google, Figure 3.9(a) shows the variations of the stock's price. The

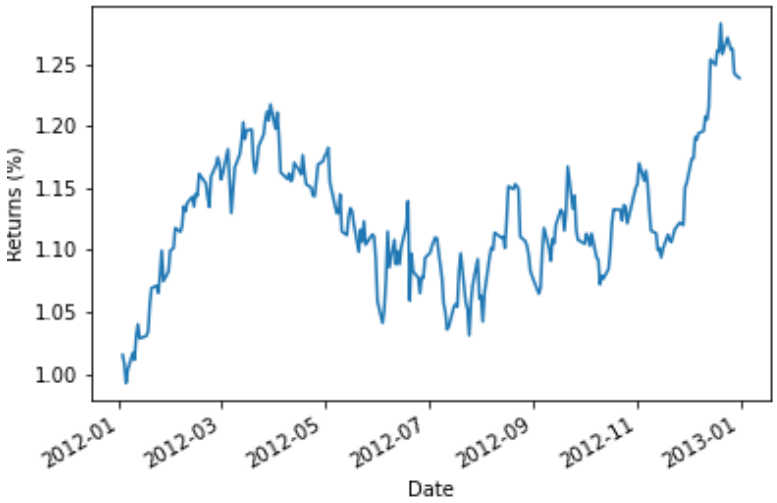
algorithm will detect the various trends and perform either a close or a sell order accordingly. Figure 3.9(b) demonstrates that, after the investment, the position is closed based on the decline but is opened after an emerging incline. At the end, a major drop also triggers the system before it orders a new buy call. The strategy tries to reduce the losses obtained while trying to maximize the profits.

This type of strategy is an example of an active investment approach, described in 2.1.3, that performs a lot of trades over the testing period. The number of actions performed will depend on the type of market, going from very low in a situation where the system doesn't leave the market to a very high amount in a volatile situation.

3.6.5 Strategy 5



(a) Stock Price



(b) Cumulative Returns

Figure 3.10: Buy & Hold summary.

Last but not least, the buy and hold (B&H) strategy was implemented. It is a very known strategy that defends opening a position and holding it no matter the trends of the market. This strategy is a perfect fit with the goal of this work, looking for good steady growth companies to collect the profits in the long-term.

Using the Adobe company's share (ADBE) as an example, Figure 3.10 shows an implementation of the algorithm. By not performing any trades after the initial investment, the system was able to return profits, ending up waiting for the hopeful rise of the stock's price, as shown by Figure 3.10(b).

The buy and hold strategy is an example of a passive investment strategy, performing only two trades with the buy order at the start of the testing period and a close order at the end of the year.

Chapter 4

Results

In this chapter, the results from the implemented system described in Chapter 3 are presented, simulating the investment according to the defined testing periods and instructions given by the model. First, a summary of the financial data used is presented, followed by an explanation of the evaluation metrics utilized to evaluate the performance of the implementation. Then, a description of the test scenarios that are conducted is made as well as an analysis for the results of each of the studies performed. Finally, an overview of the results obtained with a summary of the best and most important observations is presented.

4.1 Financial Data

As it was previously explained, the financial data retrieved is turned into financial ratios that are then used to train and test the implemented models. A summary of the ratios computed is displayed in Table 4.1. This information covers the period between 1 January 2009 until 1 January 2021 from many companies listed in the Standard and Poor's 500 (S&P 500) index. After filtering, the number of companies differs from year to year due to missing values or outliers, so Table 4.2 shows the number of companies for each testing year.

Table 4.1: Financial Ratios used in the implemented system.

Liquidity Ratios	**Leverage Ratios**	**Efficiency Ratios**	**Profitability Ratios**	**Market Value Ratios**
Quick Ratio	Debt-to-Equity	Asset Turnover	Return on Equity	Price-to-Earnings
Current Ratio	Interest Coverage	Inventory Turnover	Return on Assets	Price/Earnings-to-Growth
			Gross Profit Margin	Price-to-Book
			Operating Profit Margin	Earnings per Share
				Dividend Yield
				Dividend Pay-out

For the case studies designed, a sliding window mechanism is used meaning different datasets are used for training and testing each year. For training, information until three years prior to the testing

Table 4.2: Number of companies per year for testing.

Years	2012	2013	2014	2015	2016	2017	2018	2019	2020
Nº of Companies	346	365	372	391	410	423	445	456	464

year is used. Then, the model is given the data from the companies on the testing year, computing its results. Using these results, an investment strategy is performed over the following year for evaluation. An example of this mechanism can be seen in Figure 4.1.

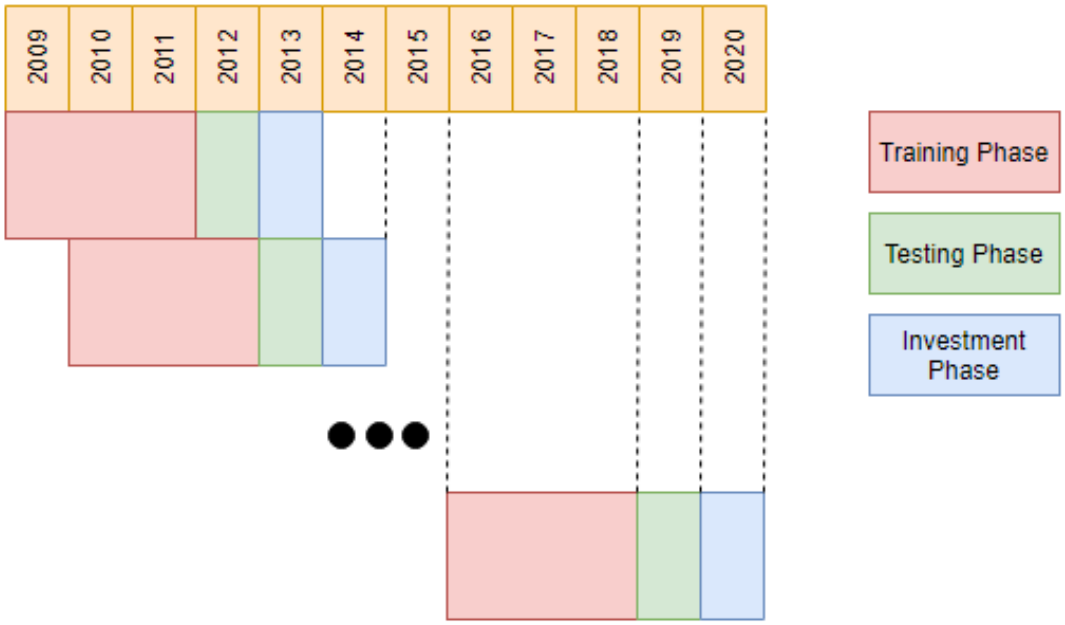


Figure 4.1: Example of the sliding window mechanism.

The testing phase corresponds to the simulation, using the models picked, where the system will compute a ranking of the companies present on the S&P 500 index in that respective year. Then, depending on the list picked, an investment simulation will be made on the following year, using the companies selected by the system, which is denominated as the investment phase.

In addition to the mechanism just described, the datasets are separated by sector as well, leading to a designed model for each sector per year. This means that, for each testing year, several different models will be created for each machine learning algorithm, which will then compute predictions that will be put together to be evaluated. More information about the S&P 500 sectors and data is given in the following subsection.

4.1.1 Sectors

The companies listed in the S&P 500 index are divided into eleven sectors based on their primary source of business. By creating a different model for each sector to be tested, restrictions have to be designed

to guarantee the legitimacy of the results obtained. In this case, ten sectors are used for testing, with Communication Services being left out due to the lack of information. A minimum number of companies threshold is also defined to make sure there is enough data for the system to be trained and tested. For this reason, only sectors with over ten companies in the index for the respective year are used, leading to the Real Estate sector being disregarded for the years 2012, 2013, 2014, and 2015. The final number of companies per sector is shown in Table 4.3. In the end, the goal is to try and maximize the effectiveness of the algorithm by establishing each model in each environment and to better understand any relationships inside each niche.

Table 4.3: Number of companies per sector per year.

Years / Sectors	2012	2013	2014	2015	2016	2017	2018	2019	2020
Consumer Discretionary	61	65	69	74	74	76	79	78	78
Consumer Staples	30	31	31	32	33	34	35	35	35
Energy	26	28	29	28	28	23	26	26	24
Financials	49	51	50	50	52	54	54	55	54
Health Care	34	38	39	43	46	49	50	51	55
Industrials	44	48	49	53	56	58	60	64	66
Information Technology	53	54	55	58	61	63	73	77	81
Materials	22	22	22	21	21	23	24	27	29
Real Estate	1	2	3	6	11	15	15	14	14
Utilities	26	26	25	26	28	28	29	29	28

4.2 Evaluation Metrics

Taking into consideration that this work is focused on a year-long investment plan with a percentage-based algorithm, the performance evaluation metrics must be focused on what better applies to this purpose. Therefore, three metrics were computed, to compare the different implementations on the tests presented in this chapter, being Return on Investment (ROI), Maximum Drawdown (MDD), and Sharpe ratio. These metrics evaluate the results in very different aspects, from being able to understand the potential in profits to measure the largest loss during the investment period and the risk-adjusted

return.

4.2.1 Return on Investment

The return on investment is one of the most used evaluation metrics in investment markets. It calculates a percentage that shows the profit, or loss, resulting from the investment made, as shown in 4.1.

$$\text{ROI [\%]} = \frac{\text{Final Capital} \quad \text{Initial Capital}}{\text{Initial Capital}} \quad 100 \tag{4.1}$$

For this work, this is the metric that will tell the potential of each model compared to a benchmark, in order to evaluate its adequacy. However, ROI does not take into consideration the risk of the portfolio's composition, meaning, for a complete evaluation, other metrics need to be combined with ROI.

4.2.2 Maximum Drawdown

Maximum Drawdown corresponds to the largest loss detected from a peak value to a trough, before the next peak, as shown in 4.2. MDD is an indicator of downside risk for the time cycle, being used to compare portfolios and indexes by measuring the possible losses of each.

$$\text{MDD [\%]} = \frac{\text{Trough Value} \quad \text{Peak Value}}{\text{Peak Value}} \quad 100 \tag{4.2}$$

It is important to note that MDD does not provide any information about the number of troughs and peaks found, nor the amount of time it takes for an investor to recover that loss if it is recovered. Due to these reasons, MDD is used together with other metrics, as a way to provide additional information.

4.2.3 Sharpe Ratio

The Sharpe ratio is used by investors as a way to evaluate the performance of a portfolio, in regards to its risk. The ratio, shown in 4.3, takes in the portfolio returns, the risk-free rate, which is the return of an investment with no risk associated, and the standard deviation, which is related to the volatility of the portfolio. This formula helps understand how much of the actual returns come from the portfolio's expected returns.

$$\text{Sharpe ratio} = \frac{R_p \quad R_f}{\sigma_p} \tag{4.3}$$

Even though the Sharpe ratio has some weaknesses, like assuming that returns are normally distributed, it is very commonly used to understand the benefits, or detriments, of adding an asset to the portfolio.

4.3 Case Studies and Objectives

Using the information and the strategy described in section 4.1, several test scenarios are created to evaluate the performance of the approach and reach the objectives defined. To do that, the case studies

have to be in accordance with the intents of the work, like testing the competitiveness of a percentage base algorithm, as is Logistic Regression, or evaluating the influence a dimensionality reduction technique does on the system.

Having this into consideration, four test scenarios were designed to analyze the overall system and all its characteristics:

1. Using the LR and SVM models to calculate a percentage of the confidence in the companies, to rank them, and perform investment strategies. With this, it is possible to evaluate and compare the two models in the same conditions, in the stock market.

2. Adding a Principal Component Analysis, to reduce the dimensionality of the data, to the system to understand the influence of this technique in the two models designed, and analyze the best way to improve results.

3. Using a different percentage threshold for the selection of the companies, it is possible to test the effect of this parameter in the application of these models to properly select the best firms to invest in.

4. Using a strategy ensemble technique, combining all the strategies in one, in a way to verify if this technique can improve results compared to singular approaches.

These scenarios were created to properly evaluate the different characteristics of this work, beginning with a simple algorithm comparison, then adding a technique to understand its influence in the system, then testing the effect of certain parameters in the outcome of results, and finally, trying to find a new way to look at investment strategies, by combining several ones.

4.4 Case Study 1 - LR Model vs SVM Model

The first case study is used to compare the two algorithms in their base forms. The models are trained with data issued three years before the testing information, which varies from 2013 to 2020, the last complete year at the time of writing. Taking the testing year into consideration, the scores obtained are then evaluated the following year, with 2021 only having available data until the beginning of September, where the investment ends. The investment is made according to the strategies already described, which will also provide an idea of the evolution of each portfolio depending on the results.

It is important to provide a bit of insight into the portfolio composition before the evaluation:

- While the top 20 list contains only a small fraction of the companies tested, the percentage one can represent a lot more depending on the confidence level of the model. The Logistic Regression algorithm consistently picked over 150 companies each year to invest, something more adequate to a hedge fund with more firepower. On the contrary, the SVM models were more selective, averaging about 60 companies over the years with 2020 being the year where it picked more companies at 90.

- Regarding the sectors chosen, the LR model primarily focused on the sectors of Health Care and Information Technology over the years while the SVM models diversified more, having Health Care as the main one but choosing different ones over the years. In the end, and even though the percentage lists represent a big portion of the sectors, the Energy sector was the least represented one in the portfolios computed.

The portfolios computed are then evaluated according to the metrics described in section 4.2, and compared with the S&P 500 index to have a benchmark that is representative of the market. Tables 4.4 and 4.5 show the annual returns for the portfolios picked by the LR model with the market annual returns, which will appear in some of the result tables for comparison and availability purposes.

Table 4.4: Return on Investment (%) for LR model with Top 20 companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	16,03	8,10	-4,55	-5,83	22,68	-3,71	10,98	20,72	16,55
Stoploss	14,75	15,78	-3,37	-2,44	32,47	2,48	25,66	-12,37	18,29
StoplossReentry	18,24	18,25	-7,29	-3,29	32,37	0,24	26,04	24,14	21,40
Momentum	16,80	2,94	-9,68	-7,16	28,16	0,64	16,30	25,47	15,48
Buy & Hold	19,96	19,95	-5,08	-3,04	32,27	0,01	27,51	31,85	23,40
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.5: Return on Investment (%) for LR model with above 60 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	32,95	8,77	-0,63	2,66	20,21	-2,80	15,43	6,32	14,96
Stoploss	32,99	14,17	2,07	1,49	25,33	-1,03	30,24	-13,26	16,56
StoplossReentry	34,72	15,10	0,44	7,07	26,32	-2,54	32,14	19,09	18,40
Momentum	25.61	6,75	-2,36	1,88	22,97	-5,40	21,39	17,53	12,44
Buy & Hold	35,28	16,29	0,76	9,71	26,18	-2,85	33,07	21,10	19,82
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

By looking at the results obtained, a few observations can be made:

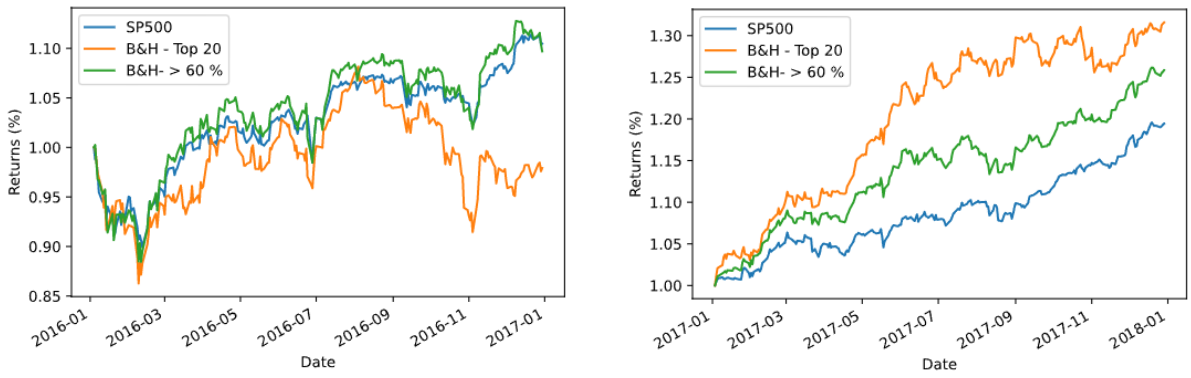
- The LR models selecting the top 20 companies shows the capability of making profits, only showing losses at the end of 2 of the 9 investment years. When compared to the S&P500 results, the model

outperforms the index in 5 of the 9 years, including staying even in a year where the index was down over 6 %.

- The percentage list case improved results, only showing losses across the board on one occasion, in 2018, and outperforming the S&P500 in 7 out of the 9 investment years. In this case, the diversification of the portfolio, with the integration of more companies, was an improvement, being able to hold off the losses in bad years and show prowess in picking the right companies in good years.

- In both lists, it is possible to conclude that the Buy & Hold strategy implemented was the one that provided higher profits, only underperforming in the years that the portfolio showed losses. Both conclusions are coherent with the work done, since the implementations were made to find outperforming long-term companies, proving that the Logistic Regression algorithm can indeed accomplish that goal. On the other hand, other strategies are more designed to prevent or smooth any losses that might occur, leading to better results in certain years.

Going deeper into the variations shown by the results of the models, it is important to understand that, since the training data used correspond to the 3 years before the testing one, any volatility showed by the sectors over the years, is going to be picked up by the models, which will influence the companies picked, and consequently, the results obtained. With this being said, Figures 4.2(a) and 4.2(b) show two cases of exactly that.



(a) 2016 (b) 2017

Figure 4.2: Cumulative Returns for Buy & Hold strategy.

In 2016, the top 20 list consisted of companies from the Health Care sector, which had been outperforming the market for the preceding 3 years but ended up presenting losses being the only sector to do so for the year of 2016. The percentage list, having a more diversified portfolio, with companies from the Information Technology, Consumer Discretionary, and Financials sector, was able to make up the losses and end up similar to the index. These sectors managed to continue their growth, outperforming the index in the prior 3 years but by much less when compared to the Health Care sector. Also important to note is that, the highest profitable sector in 2016 was Energy, but since it had shown huge losses in the prior 2 years to the testing data, it is understandable that the model ended up going in another direction.

In 2017, the opposite happened. With the previous year as the testing year, the model steered away from the Health sector, and computed a more diversified portfolio, with companies from areas that had been consistent in the prior years. Both lists ended up outperforming the index, with the Technology sector as the main focus, being the highest profitable sector in the year.

Regarding the SVM models, the results obtained, using the same process as the ones used with Logistic Regression, are shown in Tables 4.6 and 4.7.

Table 4.6: Return on Investment (%) for SVM model with top 20 companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	34,61	10,74	-3,62	0,97	18,97	-3,34	14,80	4,25	20,90
Stoploss	38,86	14,69	-1,80	-2,80	24,74	4,43	38,77	-7,92	22,51
StoplossReentry	40,28	13,49	-2,33	3,95	24,74	3,80	38,29	15,02	24,87
Momentum	27,08	11,46	-3,71	-4,90	20,60	0,09	25,44	13,90	19,87
Buy & Hold	40,37	13,11	-0,22	-4,60	24,74	2,49	39,69	16,91	26,20
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.7: Return on Investment (%) for SVM model with above 60 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	29,69	9,36	-2,20	-2,00	16,58	-1,95	16,60	12,23	14,77
Stoploss	34,03	10,90	0,94	-0,64	21,31	2,01	36,60	-8,40	17,29
StoplossReentry	36,38	10,91	1,31	4,80	21,93	1,62	37,43	26,28	18,89
Momentum	26,84	8,42	-1,91	-5,02	19,45	-1,39	26,23	23,06	13,66
Buy & Hold	36,97	12,19	3,39	5,65	21,35	-0,06	37,93	26,70	19,61
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

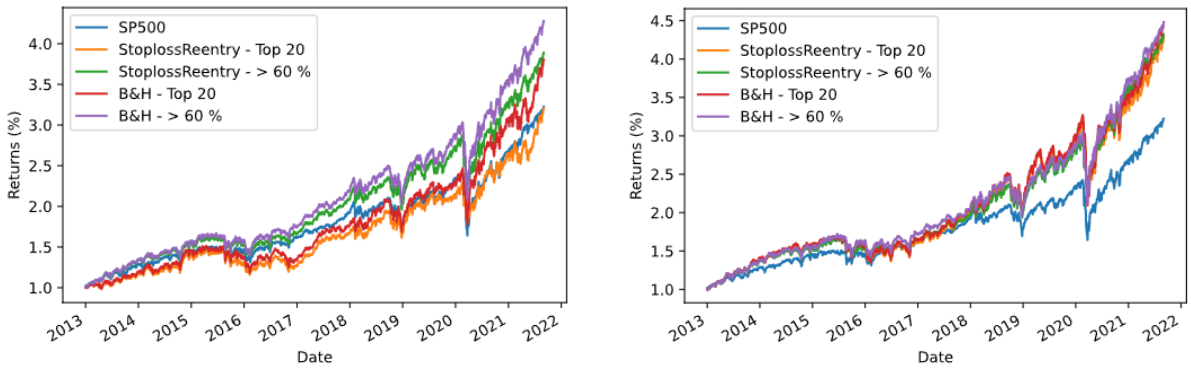
By analyzing the results, a few observations can be made:

- The differences in results between the two lists made are not as significant as the differences in the LR model. This is due to the number of companies that the LR models picked when compared to the lower quantity of the SVM models, as well as the diversification of the portfolios computed with the latter model investing in more sectors than the former.
- Both lists were able to present good profitable years, with the top 20 list only being outperformed by the index on 2 occasions, corresponding to the years where the portfolio picked showed losses.

Similarly, the percentage list was able to outperform the S&P500 index in 6 of the 9 years but was able to show profits in all of the years, depending on the strategy picked. This is due to the portfolios computed having ended the year even or profiting.

- Like the LR model, the Buy & Hold strategy was the highest performing strategy, once again confirming the research made, and outperforming the market in multiple moments over the testing period.

Finally, taking the results obtained from all the tests, a comparison is made to see the evolution of investing according to the models made in this work, when compared to the evolution of the S&P500 index over the entire investment period. With this in mind, Figures 4.3(a) and 4.3(b) show the best strategies over the years for the LR and SVM models, respectively.



(a) LR model

(b) SVM model

Figure 4.3: Cumulative returns using LR and SVM models from 2013 to 2021.

In conclusion, both the models showed capabilities in outperforming the market with the use of probability classification, with the 4 best strategies of each model being able to reach such achievement. The SVM models were able to provide better results, with the 4 best strategies outperforming all the LR ones.

4.5 Case Study 2 - Principal Component Analysis

In this case study, a PCA is added to reduce the dimensionality of the data and to test the influence a method like this has on the LR and SVM models, using probability classification. The system uses 16 ratios as features to be used by the models for predictions. The PCA computes principal components that explain the maximum amount of variance while keeping as much information as possible. The amount of variance is defined by the user, and, in this case, the value of 90 % was picked, representing a good amount of the information but reducing the number of variables considerably. In the end, the addition of the PCA reduced the 16 features to 9 or 10 principal components, depending on the sector.

First, some brief observations across the portfolios computed regarding the addition of the method:

- The amount of companies picked by both models remained the same when compared with the

system without PCA. The LR models continued to pick a big portion of the companies while the SVM continued to be more selective.

- On the contrary, the LR model with PCA showed way more diversification across the portfolio than the system without PCA. The number of sectors picked increased, with Health Care still being mostly picked but having less dominance in terms of the number of companies. Regarding the SVM models, the diversity of the portfolio continued with the addition of the PCA, reducing the weight of the Health sector across the years.

Performing the same process of testing and investment, the results for the LR models with PCA were obtained and are presented in Tables 4.8 and 4.9.

Table 4.8: Return on Investment (%) for LR model with PCA with top 20 companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	41,17	3,58	3,09	-0,29	21,37	-7,80	13,19	-9,97	13,32
Stoploss	47,47	10,67	13,66	0,22	25,58	-9,43	26,01	-12,89	15,36
StoplossReentry	48,39	11,56	12,50	7,44	25,48	-10,73	28,06	-4,20	18,47
Momentum	28,90	3,77	5,46	1,77	24,34	-10,64	18,63	-10,13	11,19
Buy & Hold	48,98	12,59	14,66	8,92	25,38	-10,50	28,52	-0,88	20,39
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.9: Return on Investment (%) for LR model with PCA with above 60 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	29,92	9,68	-0,73	0,11	16,39	-2,91	26,53	6,50	14,82
Stoploss	30,89	15,80	2,40	-3,24	19,68	-2,49	27,19	-11,62	16,54
StoplossReentry	31,99	17,14	-0,16	2,51	20,61	-4,98	29,12	16,94	18,32
Momentum	23,31	8,55	-3,34	-0,76	18,48	-5,76	18,10	15,71	12,70
Buy & Hold	32,45	18,10	0,22	5,68	20,24	-5,40	30,05	19,38	19,52
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

When comparing the results between the models with and without PCA, it is noticeable right away the fact that the addition of the PCA, resulted in the portfolio with the top 20 companies achieving great profits for the years of 2015 and 2016, years that the initial model had presented losses. However, the

system with PCA only managed to outperform its counterpart in 4 of the 9 years, including losing by 30 % in the year of 2020, since the focus of the portfolio went into the Real Estate, Utilities and Financials sectors, which were not able to recover from the pandemic as well as others.

Regarding the above 60 % list of companies, the results obtained did not show any improvements when compared to the system without PCA, being outperformed in 8 out of the 9 investment years. The results obtained showed also the same kind of variations throughout the years, being able to keep up with the market but not achieving good profitable results unlike the system without PCA.

Table 4.10: Return on Investment (%) for SVM model with PCA with top 20 companies.

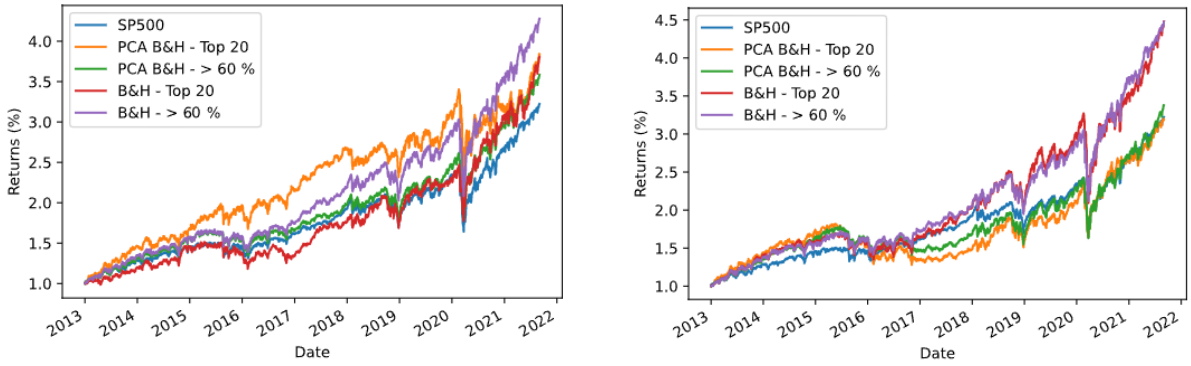
Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	40.02	15,52	-7,20	-8,08	4,58	4,38	9,08	8,73	14,00
Stoploss	44,22	20,92	-8,47	-8,66	10,00	6,94	28,92	-10,86	16,70
StoplossReentry	46,41	19,34	-9,84	-14,07	10,74	6,93	31,07	15,80	18,20
Momentum	23.30	15,37	-6,32	-15,57	7,46	6,95	19,78	16,86	12,53
Buy & Hold	46.84	19,72	-10,12	-13,58	10,32	6,52	31,23	25,97	19,20
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.11: Return on Investment (%) for SVM model with PCA with above 60 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	35,16	14,61	-3,75	-5,90	10,19	-1,27	15,30	6,29	18,19
Stoploss	36,28	21,17	-0,49	-7,18	13,95	0,20	36,46	-13,39	19,10
StoplossReentry	36,85	20,07	-1,12	-11,33	14,37	-1,89	37,39	17,46	22,04
Momentum	24,13	15,01	-1,40	-12,11	11,33	-3,42	24,85	16,17	14,73
Buy & Hold	37,27	20,04	-0,57	-10,11	14,66	-2,42	37,72	17,62	23,46
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Looking at Tables 4.10 and 4.11, it is easy to note that the addition of the PCA into the system did not produce any desired results with the SVM models. Even though the system with PCA was able to outperform the system without PCA on several occasions (4 using the top 20 list and 3 using the percentage one), the excess returns are slim when compared to the losses made in the remaining years.

To confirm this idea, a comparison was made, which is shown in Figures 4.4(a) and 4.4(b), using



(a) LR model

(b) SVM model

Figure 4.4: Comparison of cumulative returns between systems with and without PCA.

the returns obtained throughout the years, to properly see the evolution of the investment. The plots obtained confirm the conclusion regarding the SVM models, where the addition of the PCA provoked a decline in the returns, even though they were able to remain side by side with the S&P500 index. In the LR case, the PCA slightly improved the results when using the top 20 companies but worsened the profits gained considerably when using the percentage list. Also important to note that the LR models with PCA achieved better results than the SVM models with PCA, even though it is far off the base SVM models designed, as concluded in the previous section.

4.6 Case Study 3 - Percentage threshold

Using the probabilities given by the models, to classify the companies and rank them in terms of confidence, it is important to test the potential of the percentage parameter, so that the selected companies originate better results. With this being said, for this case study, the percentage list now contains only companies that are classified, by the model, as having a higher than 75 % probability of growing more than the market index.

Testing on both algorithms with and without the PCA added, some observations can be made when comparing the portfolio compositions with the tests made before:

- As expected, the change in the percentage parameter reduced considerably the number of companies picked by both systems using the LR algorithm. The average of 150 selected decreased to 75, with some years reaching a value of 100 but also reaching a smaller list with 17.
- With the SVM algorithm, both systems also reduced the number of companies selected, similarly to the Logistic Regression model. However, since the SVMs system was more selective, as explained in prior sections, using the higher percentage resulted in years with 0 companies, staying out of the market, ending up with an average of firms picked at 3.
- Regarding the sectors of the companies picked, the situation is similar to the case studies already described, with the LR model still investing in a considerable amount of companies, with more emphasis on the Technology and Health sectors. On the other hand, with the number of selected

companies being so low, the sectors invested by the SVM system were more selective and with no one having the main role.

Moving on to the investment part of the study, Tables 4.12 and 4.13 contain the returns obtained when using the percentage list ranked by the LR algorithm with and without PCA, respectively.

Table 4.12: Return on Investment (%) for LR model with above 75 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	30,43	9,59	1,16	3,29	22,68	-3,38	15,93	14,05	17,42
Stoploss	35,13	15,49	3,82	0,52	32,47	3,08	33,90	-11,04	18,04
StoplossReentry	36,18	16,66	1,74	5,98	32,37	1,35	33,95	20,89	20,17
Momentum	26,85	6,77	-1,19	3,22	28,16	2,54	22,28	19,95	13,09
Buy & Hold	36,72	18,09	2,19	8,67	32,27	1,55	35,01	26,12	21,92
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.13: Return on Investment (%) for LR model with PCA with above 75 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	32,53	7,68	-1,77	4,49	21,37	-4,58	13,93	0,07	15,55
Stoploss	37,77	13,12	2,09	-0,45	25,58	-3,27	23,08	-9,19	16,71
StoplossReentry	38,58	14,90	0,99	11,16	25,48	-5,99	27,35	2,68	18,41
Momentum	24,97	6,69	-2,52	4,16	24,34	-6,33	19,68	-2,35	11,03
Buy & Hold	38,94	17,12	1,60	11,88	25,38	-6,77	28,81	7,57	20,09
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Analyzing the tables shows that the base LR algorithm without PCA achieved massive results, only losing to the S&P500 index in the year 2016, by a slim margin. The model was capable to generate high returning years in addition to showing profit in the only year that the index showed losses.

In the year 2018, the model chose 17 companies from 4 different sectors: Information Technology, Real Estate, Health Care, and Consumer Staples. From these sectors, only Consumer Staples had bigger losses overall than the index. Looking at the portfolio, the maximum loss by a company was 2 % while the biggest profit was 5 %. Even though the S&P500 showed a decline of almost 7 %, the LR model showed an ability to pick the right companies and end the year with little profits. And in a year where the market showed such a loss, it is a good sign and achievement to stay even.

Looking at the system with PCA, it is safe to say that the percentage change did not provoke many differences, achieving similar returns to the portfolio with companies with a confidence level of 60 % or higher. However, the results obtained were still considerably better than the index growth, outperforming the market in 6 of the 9 investment years.

Table 4.14: Return on Investment (%) for SVM model with above 75 % companies.

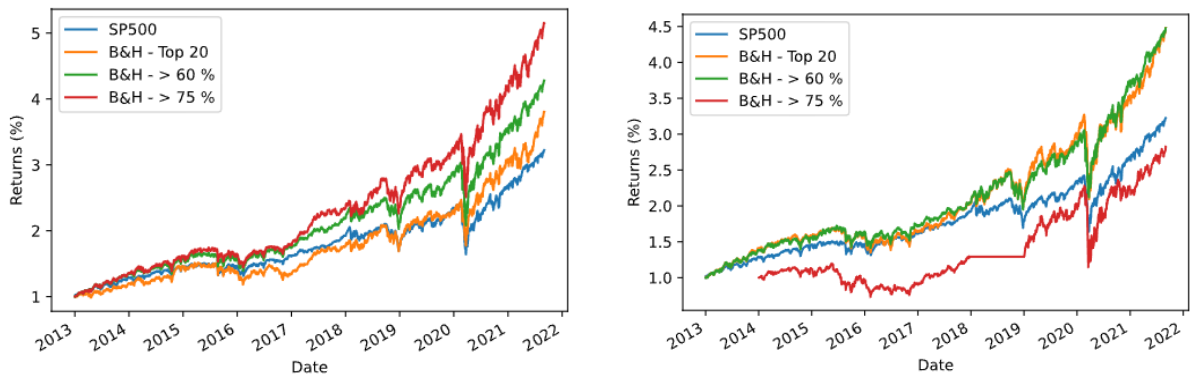
Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	–	15,73	-6,66	-8,01	39,33	–	17,81	-0,78	19,27
Stoploss	–	16,35	-15,70	-15,72	43,61	–	51,74	-23,94	26,16
StoplossReentry	–	14,38	-20,73	-10,58	43,61	–	51,74	11,58	25,17
Momentum	–	14,95	-14,12	-20,18	34,55	–	33,96	30,53	15,76
Buy & Hold	–	13,99	-15,20	-7,15	43,61	–	51,74	10,27	26,04
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.15: Return on Investment (%) for SVM model with PCA with above 75 % companies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
MA200	–	17,52	-4,92	-6,54	11,34	5,11	-16,66	24,99	–
Stoploss	–	23,71	-11,31	-8,92	10,99	-6,86	7,02	4,34	–
StoplossReentry	–	23,71	-15,05	-8,78	19,09	0,63	7,02	27,18	–
Momentum	–	13,04	-15,45	-5,30	12,20	1,46	-2,96	32.83	–
Buy & Hold	–	23,71	-11,21	-7,82	10,36	6,55	7,02	43,93	–
S&P500	29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

In the SVM case, results had mixed variations across the years, as shown in Tables 4.14 and 4.15. The system without PCA achieved amazing high return years in 2017 and 2019, but showed a lack of success in others, not being able to select the right companies. Due to the lower probability values computed by the SVM models, the system did not select any companies for the testing years of 2012 and 2017, ending up not investing in the respectively following years. The system with PCA showed similar values of companies' probability evaluations, dictating that the system stayed out of the market for the years 2013 and 2021. In the remaining years, the results obtained were more consistent, producing lower variations, which lead to lower return years.

Figures 4.5(a) and 4.5(b) shows the cumulative returns of the best results obtained in the test sce-



(a) LR model

(b) SVM model

Figure 4.5: Comparison of cumulative returns with new percentage threshold.

nario, compared to the best ones obtained in the previous studies. It is possible to observe the evolution of the highest profitable strategies of the test scenario compared to the systems with the prior percentage threshold. It is easy to confirm that the threshold change produced better results in the LR system, with the base model producing an incredible result, being the highest returning method of the research. On the SVM side, it is hard to reach a conclusion since, in both systems, there were 2 years that the system decided to stay out of investing. However, the cumulative returns of both systems did not reach the level of the S&P500 leading to the conclusion that the percentage threshold picked was inadequate to the SVM models selectivity.

These differences in the results obtained are mostly due to the probability calibration feature of each model. While both models use Platt scaling for the calibration of the probabilities, the LR model calculates them naturally while the SVM model converts the output into a probability. This should not influence the stocks' ranking in terms of confidence level but should influence the probability value of each one, where the LR algorithm shows a wider spectrum than the SVM one.

4.7 Case Study 4 - Strategy Ensemble

The final case study implemented consisted of combining all the investment strategies into one, by computing a voting system to decide which days to invest and which days to sell. Even though the best-expected strategy for this research was the Buy & Hold one, this ensemble has the goal of trying to combine what is best of every strategy, trying to obtain the high return years of finding good growing companies, but trying to cut the losses by listening to the more safe approaches as well. Having 5 strategies implemented, the voting would never end in a tie, and with the B&H as one of the implemented, the tendency would always be to hold onto the investments for the long-term. With this being said, Tables 4.16 and 4.17 present the results obtained by the 2 models, for all the systems explained in the previous studies.

Looking at the results obtained, it is fair to say that it did not work as expected. The strategy was able to maintain high return years due to the consistency of the companies picked, with most of the approaches agreeing to hold for the long-term. However, in years where it is not as linear, each strategy

Table 4.16: Return on Investment (%) for LR model using strategy ensemble.

Models \ Years		2013	2014	2015	2016	2017	2018	2019	2020	2021
No PCA	**Top 20**	16,07	15,93	-7,48	-4,40	32,37	-0,02	26,91	19,08	20,48
	> 60 %	34,11	14,66	0,24	4,38	26,09	-2,99	31,74	11,73	18,17
	> 75 %	36,00	16,07	1,80	3,86	32,37	0,15	33,97	15,55	19,77
PCA	**Top 20**	47,51	11,13	12,46	4,46	25,48	-11,10	28,77	-14,50	17,55
	> 60 %	31,47	16,69	0,07	-0,36	20,54	-4,77	28,60	10,24	18,03
	> 75 %	37,75	14,46	0,68	6,70	25,48	-5,36	28,24	-6,09	18,07
S&P500		29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

Table 4.17: Return on Investment (%) for SVM model using strategy ensemble.

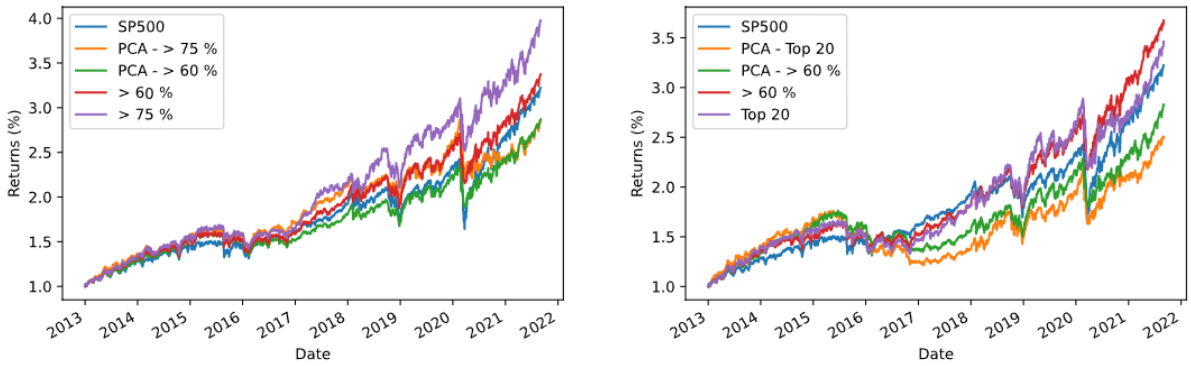
Models \ Years		2013	2014	2015	2016	2017	2018	2019	2020	2021
No PCA	**Top 20**	39,34	13,38	-2,83	-1,35	24,74	3,43	38,05	6,20	25,03
	> 60 %	35,57	10,31	0,69	0,86	21,82	1,29	37,08	19,72	18,76
	> 75 %	–	15,07	-22,56	-19,86	43,61	–	51,74	4,16	25,56
PCA	**Top 20**	19,78	19,78	-9,64	-16,82	10,77	6,97	30,11	9,50	18,20
	> 60 %	36,32	20,59	-1,14	-13,57	14,32	-1,02	37,04	9,58	21,65
	> 75 %	–	23,71	-15,21	-12,42	20,21	0,63	7,02	23,96	–
S&P500		29,20	12,80	0,08	10,47	19,43	-6,89	29,80	15,06	20,45

selected different periods to buy and sell according to their characteristics, not being able to reach a final correct decision.

The cumulative returns, shown in Figures 4.6(a) and 4.6(b), confirm the conclusions taken previously. Even though both models without PCA were able to beat the growth of the S&P500 index over the years, using different selections of companies, the results obtained do not show the same consistency, as shown in the previous case studies, nor show the ability to reach the values achieved by other strategies.

4.8 Final Discussion

Having analyzed all the case studies presented, the 3 strategies that presented better cumulative ROI, over the years, were selected for a more detailed analysis.



(a) LR model (b) SVM model

Figure 4.6: Comparison of cumulative returns using the strategy ensemble.

In terms of ROI comparison, the LR model presented in case study 3, with the percentage list limited to companies with a confidence level over 75 %, was able to achieve the highest returns of the research, returning 5,15 dollars with 1 dollar invested. Even though the SVM model was not able to reach such value, both the implementations, using the top 20 companies and the percentage list with a 60 % threshold, achieved a 4,48 on the dollar return, which is still significantly better than the 3,2 of the S&P 500 index over the investment period. Figure 4.7 shows this exact comparison.

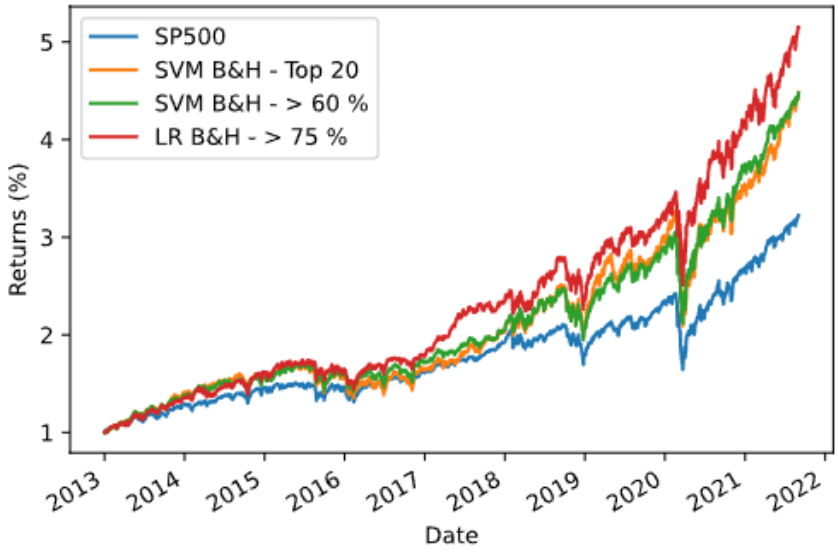


Figure 4.7: Comparison of the best strategies.

Even though the ROI metric is one of the most important metrics to evaluate the credibility of an algorithm, it is also important to verify how the results are obtained so that an investor can understand if the companies given are good in terms of potential profits but in terms of risk as well. For this reason, the Max Drawdown and the Sharpe ratio are computed to each portfolio, to measure the volatility of the portfolio and the risk associated with the investments made. Tables 4.18 and 4.19 present these measurements.

When it comes to Max Drawdown, there is no generalized threshold for distinguishing good and bad

values, but it is important to know that a 20 % drawdown value means that there was a 20 % loss swing at some point, indicating a volatile period for the portfolio, which is always undesirable. On the other hand, the Sharpe ratio has better-defined values for what investors are looking for. A value of 1.0 means there is an acceptable risk-profit relation while a ratio under 1 is sub-optimal and the profits obtained carry more risk associated than what is desired.

Table 4.18: Max Drawdown (%) comparison of the best strategies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
SVM B&H Top 20	5,63	11,36	17,54	14,97	5,87	20,94	12,18	36,04	4,84
SVM B&H 60 %	6,17	9,58	17,25	13,23	5,06	21,30	6,90	30,63	4,46
LR B&H 75 %	5,12	10,04	12,22	11,71	4,15	19,05	6,10	27,57	7,83
S&P500	5,50	7,79	12,09	10,05	2,74	19,53	6,80	32,24	3,71

Table 4.19: Sharpe ratio comparison of the best strategies.

Years / Strategies	2013	2014	2015	2016	2017	2018	2019	2020	2021
SVM B&H Top 20	2,80	0,85	0,04	0,21	2,00	0,22	1,99	0,64	2,74
SVM B&H 60 %	2,62	0,89	0,29	0,33	2,13	0,09	2,39	0,99	2,60
LR B&H 75 %	2,79	1,26	0,07	0,64	3,17	0,13	2,15	1,02	1,95
S&P500	2,43	1,09	0,01	0,82	2,89	-0,41	2,37	0,65	2,44

The measurements shown above add to the picture already painted before. The LR model using a percentage threshold of 75 % achieved, not only the highest returns of the research but showed drawdown values similar to the market and better risk-profit relations than the index, confirming exactly what this research aimed at. In regards to the SVM models, both models showed more volatility than the index, although the differences are not very meaningful. The annual Sharpe ratio calculated for the portfolios of these models also showed similar results, with the index showing slightly better values than the portfolios picked, but, with such a difference in the cumulative returns shown, the portfolios picked by both SVM models designed would still be a better option than an investment in the S&P 500 index.

Chapter 5

Conclusions

This research proposes a system that implements a Logistic Regression algorithm with the goal of maximizing the returns while investing in the stock market. The system uses a set of financial features, to capture an idea of a company's performance, and then, the LR model uses this information to compute a ranking of the companies, which will then be selected for investment. This approach is tested over 8 years using only companies included in the S&P 500 index, and with different variations, to fully conclude on the potential of this algorithm. To further evaluate the algorithm, the approach is compared to an SVM system designed as well as a market benchmark.

The first case study consisted of a simple test to compare the base LR and SVM models. Using the top 20 companies ranked by the models, and a selection of the companies ranked with a value greater than 60 %, an investment is made and results are retrieved. These results showed that both models have the capability to outperform the market, with the base SVM model beating the LR model in terms of returns obtained. The rankings computed also managed to conclude, that the LR model provides a wider spectrum of percentage values than the SVM one, leading to bigger variations in the companies selected.

These results were then used to investigate the use of a PCA on the system, to reduce the data's dimensionality. The PCA method is capable of reducing the complexity of the data while keeping most of the information intact. However, the results obtained showed less capability of the models in ranking the companies well, leading to lower returns than the ones presented in the first test scenario.

The third case study was designed to look further into the percentage parameter that ranks the companies, according to the models. Increasing the threshold from 60 % to 75 %, the investment is made only on companies that the models feel secure about. With this being said, the proposed system achieved the highest returns of the research, showing high capacity in grading the potential of a company.

The last case study was more turned into the investment strategies applied. Creating an ensemble of the investment approaches, the results obtained were not satisfactory, not reaching near the level of the Buy and Hold strategy, which was concluded to be the best strategy for this system, as expected.

Finally, the highest returning systems were compared in terms of risk associated and volatility with the market, to complete the evaluation made and understand the way the returns are obtained. These results confirmed the superiority of the LR system, not only in returns but as well in showing that the

profits were obtained more safely, and with less volatility over the years.

In conclusion, the system designed showed robust results and a high capacity to rank companies based on their financial information, proving that Logistic Regression can be a reliable tool for stock market predictions.

5.1 Future Work

In the future, several changes can be done in order to try to improve the results obtained and increase the capability of the algorithm shown in this research. Some of these changes include:

- Explore different sets of features, used to quantify the performance of companies, to try and identify the ones that represent the market evolution the best.;
- Test the application of a Genetic Algorithm with the purpose of performing feature selection and choose the best combination out of the features defined;
- More fitness functions can be tested, testing from classification metrics like f1-score to functions to maximize returns or the Sharpe ratio;
- Examine the difference of substituting the Grid Search algorithm for a Genetic Algorithm, to optimize the parameters of the LR models.

Bibliography

Albanis, G. and Batchelor, R. (2007). Combining heterogeneous classifiers for stock selection. *Intelligent Systems in Accounting, Finance & Management: International Journal*, 15(1-2):1–21.

Ali, S. S., Mubeen, M., and Hussain, A. (2018). Prediction of stock performance by using logistic regression model: evidence from pakistan stock exchange (psx). In *Patron of the Conference*, volume 15.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609.

Basu, S. (1983). The relationship between earnings' yield, market value and return for nyse common stocks: Further evidence. *Journal of financial economics*, 12(1):129–156.

Bellman, R. (1952). Dynamic programming. volume 6. Princeton University Press.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications*, 38(9):11261–11272.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

Faith, C. M. (2007). *Way of the Turtle*. MGH, New York.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Fama, E. F. and French, K. R. (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies*, 29(1):69–103.

Fama, E. F. and French, K. R. (2017). International tests of a five-factor asset pricing model. *Journal of financial Economics*, 123(3):441–463.

Gong, J. and Sun, S. (2009). A new approach of stock price prediction based on logistic regression model. In *2009 International Conference on New Trends in Information and Service Science*, pages 1366–1371.

Greenblatt, J. (2005). *The Little Book That Still Beats the Market*.

Gunsel, N. (2005). Financial ratios and the probabilistic prediction of bank failure in north cyprus. *Editorial Advisory Board e*, 18(2):191–200.

Han, S. and Chen, R.-C. (2007). Using svm with financial statement analysis for prediction of stocks. *Communications of the IIMA*, 7.

Heegaard, A. and Sørensen, P. B. R. (2013). *Analysis of stock performance based on fundamental indicators*. PhD thesis, Copenhagen Business School.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer.

Jottrand, M. (2005). Support vector machines for classification applied to facial expression analysis and remote sensing.

Lee, S. (2004). Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using gis. *Environmental Management*, 34(2):223–232.

Lee, S., Ryu, J.-H., and Kim, I.-S. (2007). Landslide susceptibility analysis and its verification using likelihood ratio, logistic regression, and artificial neural network models: case study of youngin, korea. *Landslides*, 4(4):327–338.

Maher, J. J. and Sen, T. K. (1997). Predicting bond ratings using neural networks: a comparison with logistic regression. *Intelligent Systems in Accounting, Finance & Management*, 6(1):59–72.

Min, J. H. and Jeong, C. (2009). A binary classification method for bankruptcy prediction. *Expert Systems with Applications*, 36(3):5256–5263.

Mubin, M., Iqbal, A., and Hussain, A. (2014). Determinant of return on assets and return on equity and its industry wise effects: Evidence from kse (karachi stock exchange). *Research Journal of Finance and Accounting*, 5(15):148–157.

Noori, R., Karbassi, A., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M., Farokhnia, A., and Gousheh, M. G. (2011). Assessment of input variables determination on the svm model performance using pca, gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology*, 401(3):177–189.

Öğüt, H., Doğanay, M. M., and Aktaş, R. (2009). Detecting stock-price manipulation in an emerging market: The case of turkey. *Expert Systems with Applications*, 36(9):11944–11949.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131.

Olson, D. and Mossman, C. (2003). Neural network forecasts of canadian stock returns using accounting ratios. *International Journal of Forecasting*, 19(3):453–465.

Pisner, D. A. and Schnyer, D. M. (2020). Chapter 6 - support vector machine. In Mechelli, A. and Vieira, S., editors, *Machine Learning*, pages 101–121. Academic Press.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10.

Porter, M. E. (1985). Competitive advantage, creating and sustaining superior performance.

Quah, T.-S. (2006). Improving returns on stock investment through neural network selection. In *Artificial Neural Networks in Finance and Manufacturing*, pages 152–164. IGI Global.

Silva, A., Neves, R., and Horta, N. (2015). A hybrid approach to portfolio composition based on fundamental and technical indicators. *Expert Systems with Applications*, 42(4):2036–2048.

Tsai, C.-F., Lin, Y.-C., Yen, D. C., and Chen, Y.-M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing*, 11(2):2452–2459.

Upadhyay, A., Bandyopadhyay, G., and Dutta, A. (2012). Forecasting stock performance in indian market using multinomial logistic regression. *Journal of Business Studies Quarterly*, 3(3):16.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:774–780.

Wang, H. and Hu, D. (2005). Comparison of svm and ls-svm for regression. In *2005 International Conference on Neural Networks and Brain*, volume 1, pages 279–283.

Xie, C., Luo, C., and Yu, X. (2011). Financial distress prediction based on svm and mda methods: the case of chinese listed companies. *Qual Quant*, 45:671–686.

Yu, L., Chen, H., Wang, S., and Lai, K. K. (2009). Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on Evolutionary Computation*, 13(1):87–102.

Zavgren, C. V. (1985). Assessing the vulnerability to failure of american industrial firms: a logistic analysis. *Journal of Business Finance & Accounting*, 12(1):19–45.