Portfolio Optimization using Fundamental Analysis with a Logistic Regression Approach (November 2021)

Francisco Rodrigues, Master's Student

***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.**

***Index Terms*—Financial Ratios, Financial Statements, Fundamental Analysis, Logistic Regression, S&P500, Support Vector Machines**

I. 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.

This work has been submitted for review on the 1^{st} of November 2021.

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.

The main contributions of this thesis are: the combination of Logistic Regression and Fundamental Analysis to the stock market prediction world. Although each topic has had its uses in several fields, the combination for a 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.

This work is organized as follows: Section II presents the necessary theory and methodologies used to perform this

investigation including financial statements, ratios, logistic regression, and support vector machines. This section also presents relevant studies about the topics mentioned. Section III describes the architecture of the system implemented and provides an explanation to all the different layers of the system. Section IV 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. Section V summarizes the conclusions of this work and discusses possible future work propositions.

II. Related Work

A. 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.

(Mubin, Iqbal and Hussain, 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.

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.

(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, Mubeen and Hussain, 2018) also attempted to predict stock performance using a combination of accounting and financial ratios, in the Pakistan Stock Exchange (PSE). Results obtained give the indication 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.

(Silva, Neves and Horta, 2015) combined both FA and TA with a Multi-Objective Evolutionary Algorithm (MOEA) 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, Lin, Yen and Chen, 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.

B. 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.

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, Doğanay and Aktaş, 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 out-performing 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, Bandyopadhyay and Dutta, 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.

C. 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, Guyon and Vapnik, 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, Luo and Yu, 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.

III. System Architecture

This work presents a combination of a Logistic Regression model with Fundamental Analysis to predict which companies will have a major increase in value with the goal of maximizing profits. 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 in 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 to compare performances. As illustrated in Figure 1, the system's architecture is divided into five layers. Each layer has its role in the system and is independent of 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:

The **Data Preparation Layer** is in charge of obtaining the data necessary for the model, filtering it, and preparing 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 to have the ratios desired so that the models created can be trained and tested.

The **Training Layer** was created to tune the models described. Since both Logistic Regression and Support Vector Machine have parameters to be optimized, an extensive search has to be made to guarantee an appropriate model for each circumstance. Ultimately, the models use the training set to evaluate many 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 for 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 behavior of each choice.

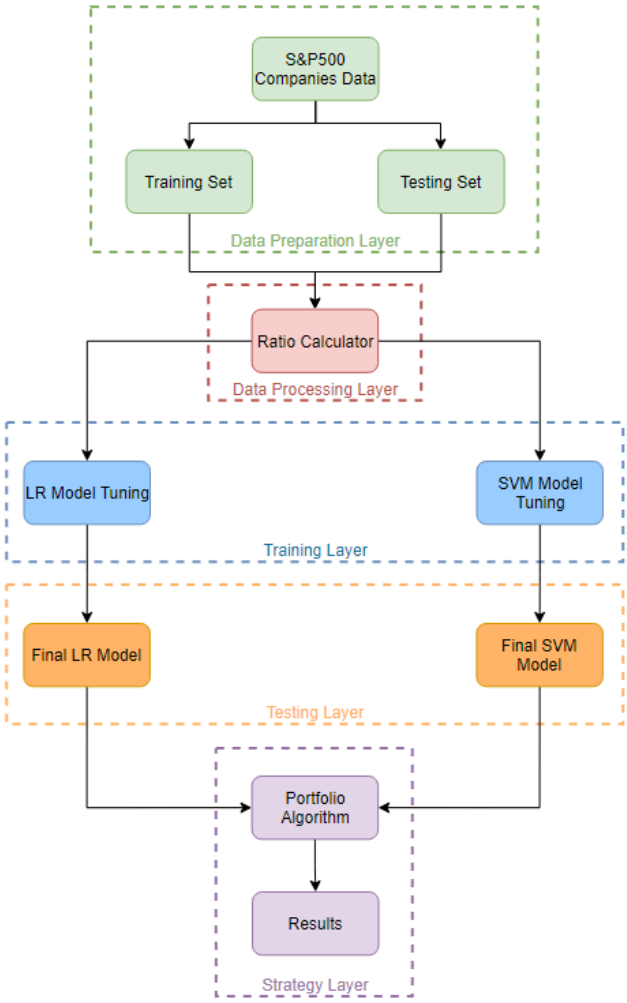

Figure 1: Flowchart of the overall system

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

A. 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 into 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 to finish 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 the companies present in each year.

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 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 with 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 expect to be 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.

B. 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 III-A1 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 computed are shown in Table I.

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.

1) 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

Table I: Financial ratios and observations

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

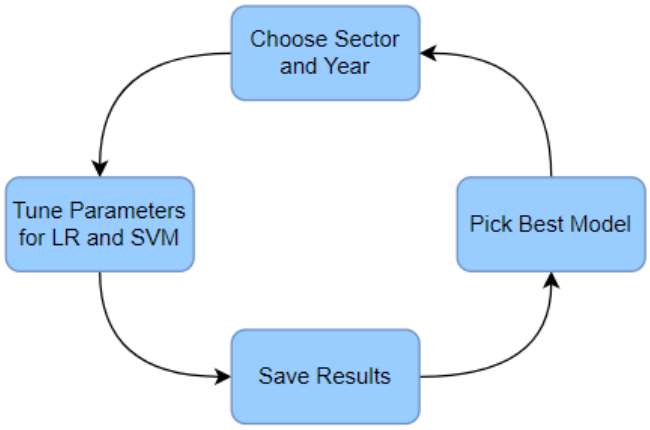

Figure 2: Architecture of the Training Layer

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 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 a real insight into the potential of the investments. Due to all these 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 to provide out-performing results and become a noticeable research for the community.

C. Training Layer

The training layer is tasked with training each model to find the one that guarantees the best results in the testing phase. First, the sector and year are selected. This is done since each industry has very different implications for 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 the 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, totaling 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 the final results.

1) Fitness Function

The fitness function used in this work was the ROC-AUC. 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.

D. 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 data frame. 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.

E. Portfolio Layer

The portfolio layer is responsible for finding a way to make money with the investments in 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 a 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 buying and close orders in the stock market. This will provide a good insight into 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.

1) Strategies

Five different investment strategies were implemented, trying to cover all the different types of markets. The first one is designated as MA200, which takes the 200-day moving average, and opens and closes positions according to the price being higher or lower than the average. The second strategy is designated as Stoploss, basically closing a position if the price of a stock lowers past a threshold. The third strategy is designated as StoplossReentry. It is very similar to the second one but tries to find a positive trend to reenter the market and profit off market recoveries. The fourth strategy is designated as Momentum. This one computes a series of market streaks, trying to find negative and positive trends to base its decision on. 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.

IV. Results

A. 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.

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 1.

$$\text{ROI } [\%] = \frac{\text{Final Capital} - \text{Initial Capital}}{\text{Initial Capital}} \times 100 \quad (1)$$

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.

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 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.

$$\text{MDD } [\%] = \frac{\text{Trough Value} - \text{Peak Value}}{\text{Peak Value}} \times 100 \quad (2)$$

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.

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 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.

$$\text{Sharpe ratio} = \frac{R_p - R_f}{\sigma_p} \quad (3)$$

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.

B. 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. Considering the testing year, 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.

The portfolios computed are then evaluated according to the Return on Investment (ROI), Max Drawdown, and Sharpe ratio metrics, and compared with the S&P 500 index to have a benchmark that is representative of the market.

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.

C. 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

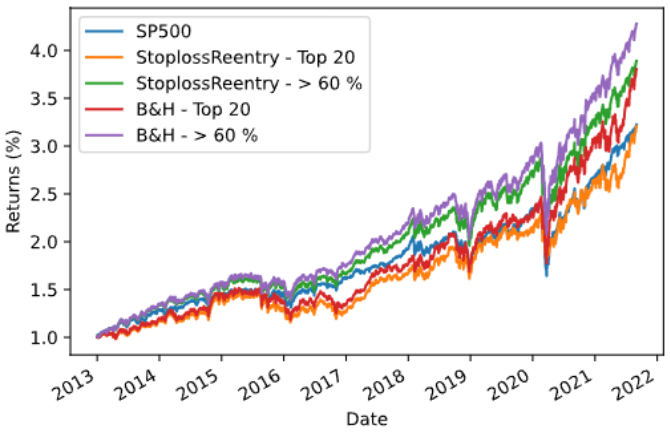

Figure 3: Cumulative returns using LR models from 2013 to 2021

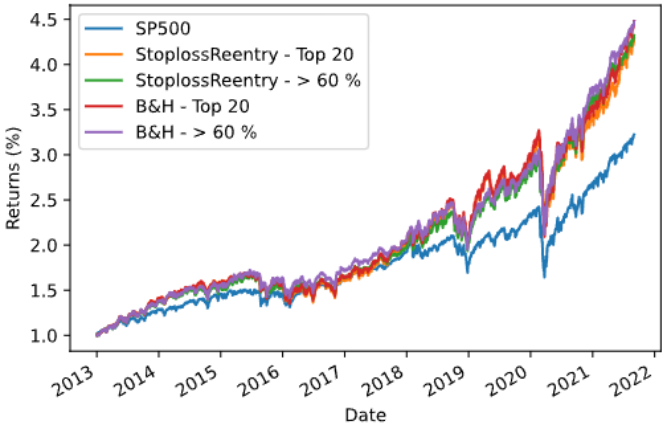

Figure 4: Cumulative returns using SVM models from 2013 to 2021

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.

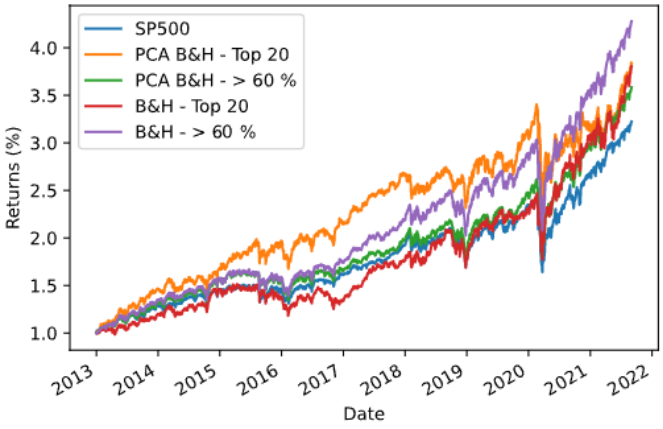

Figure 5: Comparison of cumulative returns between LR models with and without PCA

To confirm this idea, a comparison was made, which is

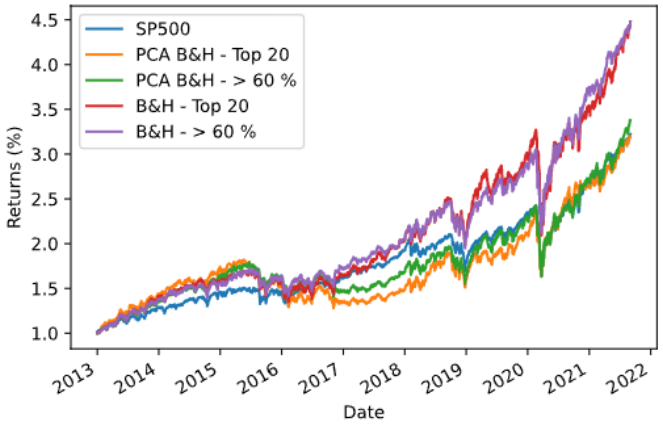

Figure 6: Comparison of cumulative returns between SVM models with and without PCA

shown in Figures 5 and 6, using 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.

D. 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.

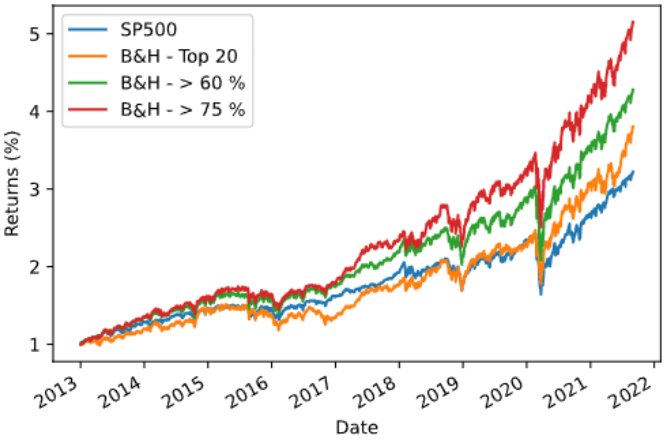

Figure 7: Comparison of cumulative returns with new percentage threshold

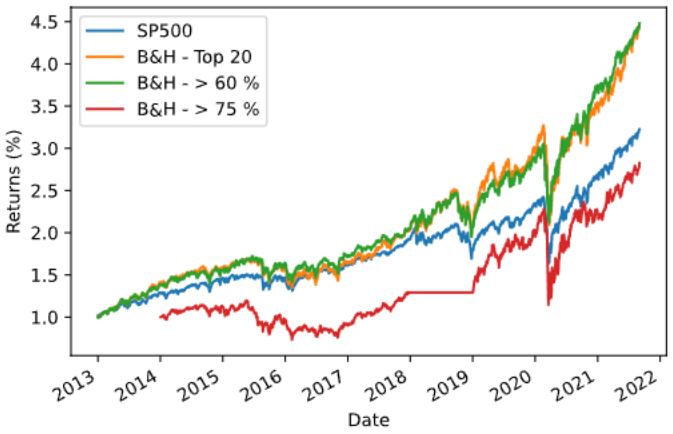

Figure 8: Comparison of cumulative returns with new percentage threshold

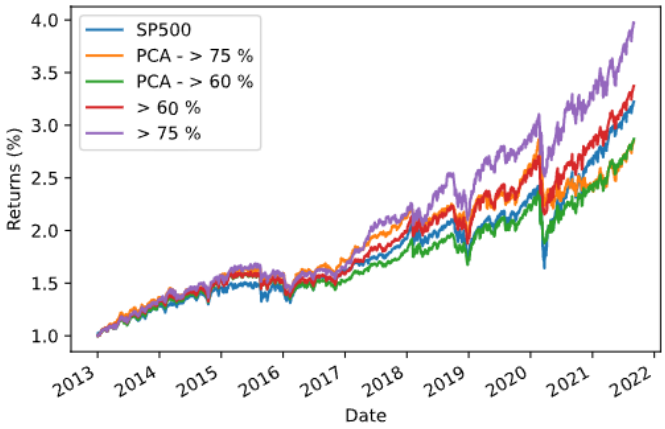

Figure 9: Comparison of cumulative returns using the strategy ensemble

Figures 7 and 8 shows the cumulative returns of the best results obtained in the test scenario, 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.

E. 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.

The cumulative returns, shown in Figures 9 and 10, 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,

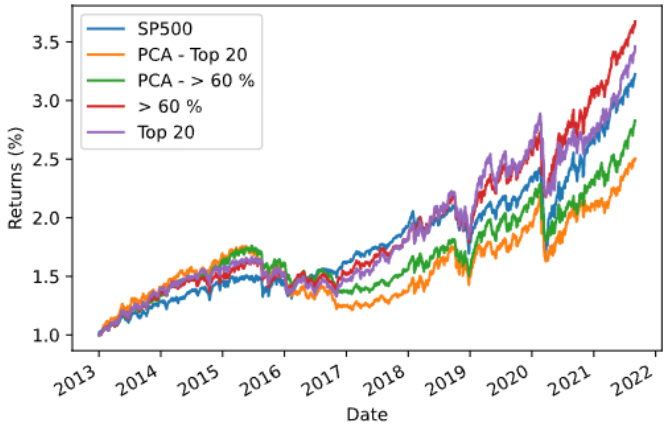

Figure 10: Comparison of cumulative returns using the strategy ensemble

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

F. 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.

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 11 shows this exact comparison.

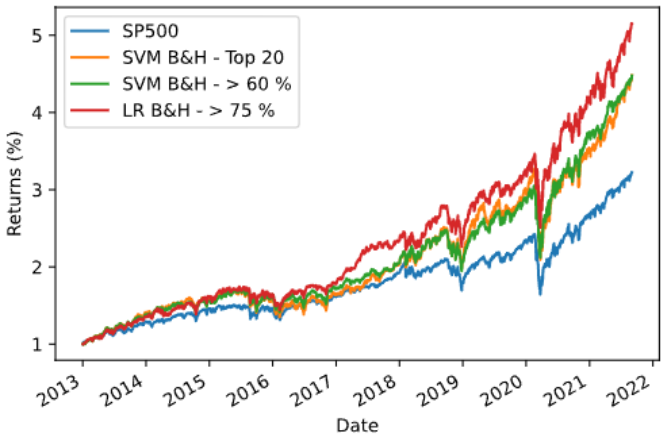

Figure 11: Comparison of the best strategies

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.

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

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

Table III: Sharpe ratio comparison of the best strategies

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

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.

V. 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, in order 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.

A. 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 with the purpose of maximizing 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.

References

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

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

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

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers, in: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Association for Computing Machinery, New York, NY, USA. p. 144–152.

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

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297.

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

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

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

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

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

Gong, J., 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, pp. 1366–1371.

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

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

Heegaard, A., Sørensen, P.B.R., 2013. Analysis of stock performance based on fundamental indicators. Ph.D. thesis. Copenhagen Business School.

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

Mubin, M., Iqbal, A., 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, 148–157.

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

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

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

Tsai, C.F., Lin, Y.C., Yen, D.C., Chen, Y.M., 2011. Predicting stock returns by classifier ensembles. Applied Soft Computing 11, 2452–2459.

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

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

Wang, H., Hu, D., 2005. Comparison of svm and ls-svm for regression, in: 2005 International Conference on Neural Networks and Brain, pp. 279–283.

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

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