uc3m | Universidad Carlos III de Madrid

Master's degree in Big Data Analytics
2022-2023

Master's Thesis

Stocks selection: a medium/long-term approach using Machine Learning techniques

Francisco Esteves Muñoz

Ricardo Aler Mur

Madrid, June the 30th, 2023

SUMMARY

Stock selection is the act of identifying and choosing specific stocks to include in a portfolio based on certain criteria to match investors objectives. In this study, we propose a machine learning (ML) technique for approaching medium to long term investments using big market capitalization companies from the US, that is able to consistently outperform the market. The data acquired for the ML model construction, consists of fundamental data of the companies listed in the mayor US stock markets and macroeconomic data. The expanding window method is employed with a one-year time horizon, spanning from 2000 to 2021. An extremely randomized tree (ET) classification model is applied to 22 different data sets. Each year, the portfolio is selected based on the companies with higher probabilities. Subsequently, the return over that portfolio was calculated for the year after, avoiding data leakage from the future. We have also tuned the most important hyperparameters of that model and prove that there is enough statistical evidence to affirm that our model beats the market consistently, overruling the efficient market hypothesis (EMH). Finally, the variability of the model was studied for the construction of the final model, in which the trade-off between stability and higher returns was taken into consideration.

Keywords: Machine learning; Stock; Portfolio; Investment; Fundamental analysis; Macroeconomic data.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Motivation of the work

The stock market has long provided investors with opportunities to generate significant returns on their investments. However, the challenge lies in selecting an optimal portfolio that aligns with investors' objectives and desired returns. Traditional methodologies such as fundamental and technical analysis rely on manual processes, resulting in biased selection and time-consuming procedures. The advent of new technologies and the abundance of market data have prompted the exploration of alternative approaches to these methodologies. The rapid growth of artificial intelligence (AI) and machine learning (ML) has given rise to various research directions and methodologies aimed at leveraging these technologies for the benefit of investors.

While ML techniques for stock investment have been extensively studied in recent years, the focus has predominantly been on applying these techniques to technical analysis for short time periods [1], with an emphasis on seeking instant rewards. Conversely, medium to long-term investments have received less attention, primarily due to the challenges associated with gathering fundamental data on companies and the absence of "instantaneous" rewards.

Hence, the motivation of this study is to harness the power of ML and apply it to stock investments, with a specific focus on medium to long-term periods and utilizing fundamental variables of companies, along with macroeconomic data. By doing so, this study aims to not only to generate useful models for investors that provide objective selections and measurable insights but also benefit society and stock markets as a whole by facilitating optimal capital allocation.

1.2 Objectives

The primary objective of this study is to develop a robust and effective approach for stock selection using machine learning techniques specifically tailored for medium to long-term investments. To consider the model useful in this particular subject, the following requirements must be fulfilled:

- Consistently Outperforming the Market: The model should generate returns that consistently surpass the performance of the overall market and common stock market indices such as the S&P 500, NASDAQ composite, etc. This benchmark outperformance demonstrates the model's ability to identify stocks with superior growth potential or undervalued assets.
- Stability in Varying Market Conditions: The model should exhibit stability, meaning that its performance remains relatively consistent across bullish and bearish market periods. It should not be excessively influenced by short-term market fluctuations but maintain a reliable and robust performance over various market cycles.

- Feasibility for Small Investors: The outcomes and recommendations of the model should be feasible for small investors to replicate. The model should not rely on access to exclusive resources or specialized knowledge that is beyond the reach of individual investors. It should provide practical and actionable insights that can be implemented by investors with limited resources.

In addition to the primary objectives, there are secondary objectives for this study:

- Trustworthy Dataset: The study aims to gather a trustworthy and comprehensive dataset that includes not only companies' fundamental data but also macroeconomic indicators. This dataset will serve as the foundation for training and testing the machine learning models, ensuring the accuracy and reliability of the results.
- Investigation of Efficient Market Hypothesis (EMH): The study intends to explore the relationship between machine learning techniques and the efficient market hypothesis. It aims to investigate whether ML models can exploit inefficiencies or patterns in the market that may contradict the assumptions of EMH. This analysis will contribute to the understanding of how ML can enhance stock selection strategies.

By fulfilling these requirements and objectives, the study seeks to provide a valuable and practical framework for stock selection using machine learning techniques.

2. BACKGROUND

2.1 Stock investment

A stock or equity is an assurance that secure the ownership of a fraction of the corporation being issued. The amount of the stocks which someone possess, can be called "share" which entitles the owner to a proportion of the corporation's assets and profits. Stocks are interchanged mainly in stock exchanges and represent the foundation of modern investing and the pillars of investing for some of the most important invertors nowadays.

The stock market starts with the need of the companies of raising capital to expand and the investors necessity of diversifying the risks. However, for a company to be listed in a stock exchange market, there are several requirements to be fulfilled. Those requirements vary depending on the exchange, but the most important ones are to have a big enough market capitalization[1] and to have liquidity[2] of shares (for that, some shares must be issued before being listed). Some of the requirements for companies listed in the New York Stock Exchange (NYSE) and National Association of Securities Dealers Automated Quotations (NASDAQ) can be found on [2]. Apart from those requirements, companies' reports must be issued periodically and provide regular financial statements, including balance sheets, income statements, and cash flow statements.

There are two main ways in which an investor can make profits when buying a stock:

1. Capital Appreciation: when the stock value of a company increases, the investor can sell that stock to make a profit.
2. Dividends: Some companies distribute a portion of their profits to shareholders in the form of dividends. Dividends are usually paid on a regular basis (quarterly, semi-annually, or annually) and represent a cash return on investment.

The earning made by investors in the past, is growing continuously and the globalization with the introduction of new technologies, has allowed the exponential increase of market capitalization over the years. In 2020, the total market capitalization of companies around the world, was 93.69 trillion US dollars according to the World Bank [3]. From that total market capitalization, undoubtedly, the US gather the mayor stock exchanges in the world. In January 2023, US stock exchanges represented the 58.4% of the whole world market capitalization [4].

There are two main approaches to evaluate a stock investment:

Technical analysis

This technique is widely used and has many variants, but the key concept is to study historical data to predict future price movements. The whole idea behind is that stock price patterns repeat over time. The main variables when analysing a stock in this way are past prices, past trends, and market volumes. Technical analysis can be used for intra-day trading which open and close all operations during a single day and for looking at good entering prices for a medium-long term investment.

Fundamental analysis

[1] Total value of the stocks issued by a company or market.

[2] Capacity to turn an asset into cash.

Is the idea of evaluating a company by its ability to grow and create wealth. It focuses on analysing financial statements, balance sheets, economic indicators, company's new products and management. The key idea is to identify medium/long-term investments opportunities based on that if the company presents healthy reports and the price is reasonable, the stock price will increase.

For fundamental analysts, quarterly reports provide essential information. Not only they describe the future of the company, new products, developments, management strategies, but also, they show the company's numbers. Some key numbers shown on the reports are the following:

1. Net income, also referred to as earnings, is calculated as sales minus cost of goods sold, selling, general and administrative expenses, operating expenses, depreciation, interest, taxes, and other expenses. It represents the residual amount of revenue left after deducting all expenses, taxes, and interests.
2. Assets are the resources that the company owns, that have measurable economic value. Usually, it includes tangible assets such as inventories, receivables, plant and equipment, as well as intangible assets like patents, trademarks, and goodwill. Other typical assets listed are cash, investments, and accounts receivable.
3. Liabilities represent the financial obligations or debts owed by a company. Liabilities reflect what a company owes and need to be repaid or fulfilled over time. Analysts watch very carefully this number, as Peter Lynch explained in [5], "Companies that have no debt can't go bankrupt."
4. "Cash flow is the amount of money a company takes in as a result of doing business" [5]. Positive cash flow indicates that a company is generating more cash than it is spending, while negative cash flow suggests a cash outflow. Free cash flow, which is the important metric for investors, is a measure of the cash which a company generates after deducting its operating expenses and capital expenditures. It represents the cash available for distribution to investors, debt repayment, or reinvestment in the business.

From the reports and the obtained numbers, there are key ratios that analysts use, some of those are presented below:

1. Earnings per share (EPS) this ratio measure the profitability of the company:

$$EPS = \frac{Earnings - Preferred\ dividends}{weighted\ average\ number\ of\ common\ shares} \qquad (2\text{-}1)$$

Note: weighted average number of common shares is set to correct the number of shares in case the company issues new shares or repurchase them.

2. Price-earnings (P/E) ratio measures the potential growth of the stock price:

$$P/E = \frac{Price * Outstanding\ shares}{Earnings} \qquad (2\text{-}2)$$

3. Debt to equity (D/E) ratio measures the financial leverage of the company:

$$D/E = \frac{Total\ liabilities}{Total\ shareholders'equity} \qquad (2\text{-}3)$$

4. Return on equity (ROE) measures profitability and how effectively a company uses shareholder money to make a profit:

$$ROE = \frac{Earnings}{average\ shareholders'equity} \qquad (2\text{-}4)$$

Note: average shareholders' equity is set to account for the equity between the beginning and the end of the period.

5. Working capital (WCR) ratio measures how easily a company can turn assets into cash to pay short-term obligations:

$$WCR = \frac{Current\ assets}{Current\ liabilities} \qquad (2\text{-}5)$$

6. Quick ratio (QR) is another measure for liquidity but discount the inventories which takes more to convert into cash:

$$QR = \frac{Quick\ \ assets}{Current\ liabilities} \qquad (2\text{-}6)$$

Where:

- *Quick assets = CE + MS + NAR = TCA – Inventory - PE*
- *CE: cash equivalent*
- *MS: marketable securities*
- *NAR: net accounts receivable*
- *TCA: total current assets*
- *PE: prepaid expenses*

Stock splits

It occurs when companies divide its existing shares into multiple shares, for the investor the capital invested remains the same, but the stock will be traded now as a fraction of what was traded before. The reason for this, could be many but the most important ones are to avoid large stock prices (which typically repel investors) and to enhance the liquidity.

This phenomenon is not that important per se, but is key when analysing investments through the years. For calculating the true growth of a stock, it is required to adjust it by splits that may have occurred in the meantime.

Efficient Market Hypothesis

The efficient market hypothesis (EMH) states that the market already reflects past information, and due to that, all movements in prices are due to new information. Behind this theory, it is impossible to perform any analysis (technical nor fundamental) to predict if a stock will go up. The EMH is based on the idea that there are rational and profit-seeking investors who are analysing all the information on a consistent basis and quickly incorporate new information into the prices of assets.

As pointed out by Burton G. Malkiel in 2003 [5], "a generation ago, EMH was widely accepted by academic financial economists…", and it was "…generally believed that securities markets were extremely efficient in reflecting information about individual stocks and about the stock market as a whole”. However, by the start of the XXI century, the EMH started to lose ground around investors, financial economist and statisticians and begun the sense that stock prices are “at least partially predictable”.

The current situation is that the original EMH has diverged into three main theories [7]:

1. Weak form implies that the stock prices already reflect the data of the past prices and operating volumes and therefore no technical analysis can be used to outperform the market.
2. Semi-strong form states that all the past public information is already reflected in the stock prices, implying that investors cannot utilize technical or fundamental analysis. In this form, only private information could lead to an outperform of the market.
3. Strong form theory states that all information, public and not public, is completely accounted for in current stock prices. The result of this theory is that no type of information could lead to a constant outperformance of the market.

Nowadays, with the proof of many smart and strategic investors outperforming the market using different strategies, the strong and semi-strong theories have become obsolete. The weak form, however, is still being discussed between those who think technical analysis is useful to predict stock prices and those who think it is not.

US stock markets

As pointed out before, the US stock market is by far the most important, not only because of the volume traded daily but also for its tradition. The most important stock exchanges in the US are NYSE, NASDAQ and Chicago Stock Exchange (CHX). When following stock markets, many analysts use indexes to know how well the market is performing as a whole.

Standard & Poor's 500 (S&P 500) index, is a market capitalization-weighted index of the 500 leading publicly traded companies in the U.S. and is the basic index followed by the traders. Another important U.S. stock market benchmark is the Dow Jones Industrial Average (DJIA), this index consists of 30 companies associated with retail and has been widely used in the XX century. NASDAQ, which is an electronic marketplace, has also many indices that help an investor to know how well is that exchange performing, the NASDAQ composite is the preferred index when looking a tech companies.

From the indices presented before, the S&P 500 is the preferred by institutional investors, given the large amount (but not exaggerated) of companies and the diversity of sectors included. Unlike NASDAQ composite and DJIA, the S&P 500 gather companies from all industries and its weighted market capitalization average composition, makes the index more stable and representative from the market reality.

As important as market indicators and profits, is the regulatory environment. This is how to ensure fair and transparent markets, prevent fraud, and maintain market integrity. In the US, the Securities and Exchange Commission (SEC) is the main regulatory body overseeing market’s security, enforcing rules, and protecting investors.

Also, in the last decades, with the advance of technologies and data acquisition, new ways of investing have arrived. For example, the stock market has witnessed significant changes with the introduction of high-frequency trading[3] (HFT) and algorithmic trading. More recently, the arrival of big data and artificial intelligence (AI) have influenced the market into finding extremely refined models that find patterns to optimize investments.

Importance of the economy in stock markets

One thing that it is usually called out when talking about stock investments, is the volatility of prices in the market. Reasons for that are the supply and demand forces exchanging constantly stocks as prices go up and down. The excess of speculation is one of the market problems as Jana Drutarovská pointed out in [8], describing the disconnection between the financial markets and the real economy and the performance of the companies. That being said, macroeconomic policies carried out by governments, have huge impacts on investors decisions. It is very typical to see bearish[4] weeks when the Federal Reserve (FED) increases suddenly its interest rates and economists predict crisis.

Not only in sudden changes, is important to notice macroeconomic variables, but also for the long term, as those policies led to healthy or sick economies and markets. For example, in 2008 extremely low interest rates, led to a boom in the housing market (and stock market as well) generating excessive and unrealistic high prices, which after led to a boost. Because of that, S&P 500 fell almost 20% in just one week.

According to [9], macroeconomic indicators are key for investors decisions, some of the most important ones are:

- Gross domestic product (GDP) is the market value of all goods and services produced within a country during a given period.
- Interest rates are the percentage charged on loans or paid to the owners of savings accounts. The interest rate is set by the FED in the US, and have a direct impact on the capital allocation, higher interest rates allocates the capital into bonds, money market-funds, etc.
- Inflation is known as the continuous growth in the prices of an economy, it is measure as a rate. It leads into a malfunction of the economy prices signals and lowers the purchasing power of the inhabitants. It is extremely related with interest rates, and many central banks use interest mechanisms in order to control inflation rates.
- Unemployment is the portion of the labour force of those who are looking for a job but are unemployed. Unemployment rate is negatively correlated with inflation in the short/medium-term.

[3] Method of trading that uses powerful computer programs to transact a large number of orders in fractions of a second.

[4] Drastic down in the stock market, typically more than 20%.

2.2 Machine learning background

Machine learning (ML) is a fast-evolving subfield of Artificial Intelligence (AI) which develops algorithms that are designed to emulate human intelligence by predictions or decisions without being explicitly programmed to do so. Techniques based on machine learning have been applied in fields such as pattern recognition, computer vision, engineering, finance, entertainment, marketing, and biomedical and medical applications.

As introduced before, machine learning in stock investments, has been used mainly for stock prediction, portfolio optimization, algorithmic trading and fraud detection. Some techniques and models that are used in these applications and in this thesis in particular, are explained briefly below.

2.2.1 Data pre-processing

This step in ML refers to the techniques applied to transform raw data into data that can be fed to a model for obtaining an optimal performance. Usually, data comes in with missing values, outliers, irrelevant features and different formats.

Imputation with K-Nearest neighbours

This methodology provides imputation[5] for missing values using K-Nearest neighbours (*see 2.2.2*) approach. The strategy is to approximate missing values with an average (or weighted average) of the existing values of the *k* points which are closer to that instance.

2.2.2 Model evaluation

Train/Validation/Test split

This split is done in order to obtain an evaluation on how well the model would perform in unseen data. When working with time dependent data, there are two main approaches followed for making that split: expanding window and sliding window. Expanding window, uses all the training data before the testing time point, while sliding window uses a *w-hyperparameter* which controls the depth of the training set, see *Fig. 2-1.*

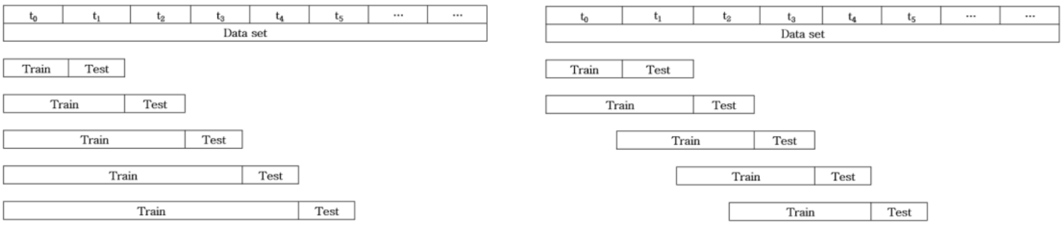

Fig. 2-1: Expanding window and sliding window.

2.2.3 Machine learning models

K-Nearest neighbours

[5] Process of replacing missing data with approximations of the true value.

K-Nearest Neighbours (KNN) is a simple and intuitive algorithm for both classification and regression. The idea behind KNN is that similar points are close in the feature space. Its advantage is the simplicity behind the algorithm, and a few numbers of hyperparameters[6] usually make it an easy first model. The value of *k* determines the model's sensitivity to local variations. The main disadvantage is that it turns computationally expensive for large datasets when making predictions.

Random forest and extremely randomized trees

Random Forest (RF) is an ensemble learning method that combines multiple decision trees to make predictions. It can be used as a regressor model and a classification model as well. Each tree is trained on a bootstrapped[7] set of the initial data, and the final prediction is made by aggregating the predictions of individual trees. The main advantages of random forest are its versatility and its ability to work with complex data. Extremely Randomized Trees, also known as Extra-Trees (ET), is another ensemble learning method similar to Random Forest, where each tree is trained on the whole data set, but at each node a subset of the variables is randomly selected. Also, the thresholds for numeric variables are set at random, decreasing the computational cost of training. The main advantage of Extra-Trees is the ability to deal with complex data while being a fast model to train.

Another important advantage of these models is that they provide a useful in-hand tool that quantifies the importance of features.

Gradient boosting

Gradient boosting is an ensemble ML technique used for both regression and classification tasks. This model combines multiple weak predictive models (often decision trees) to create a more accurate and robust predictive model. The prediction then, is obtained by aggregating all the predictions made by the weak models.

The basic idea behind gradient boosting is to iteratively train a sequence of weak models, where each model tries to correct the mistakes made by the previous models in the sequence. The models are built in a forward stage-wise manner, with each new model attempting to minimize the errors or residuals of the previous models. The main advantage of this model is that it can capture complex patterns and relationships in the data.

Support vector machines

Support Vector Machines (SVM) are supervised learning models that achieve outstanding results in both classification and regression tasks. When dealing with a classification problem, SVM key idea is to find a non-linear transformation in which the data could be optimally separated by a hyperplane, maximizing the margin between two different classes. In regression, the non-linear transformation leads to a space in which a linear regression can be optimally applied.

The kernel function calculates the similarity or inner product between pairs of data points and in the original feature space or the transformed feature space. An accurate kernel selection and tuning of the kernels parameters is key to obtain good results with these models.

[6] A variable that sets the complexity of the model, which can typically be used to set an adequate bias-variance trade off.

[7] Randomly sampling the original dataset with replacement.

3. STATE OF THE ART

3.1 Machine learning in the investment context

When explaining the new changes that markets are experiencing, it is impossible not to talk about the technology advance over the investing strategies. In this sense, there are continuous progress in fields such as data analytics, risk management, portfolio optimization, high-frequency trading, sentiment analysis and algorithmic trading.

This explosive development has been allowed by the increase in computing power over the last decades, the advancements in artificial intelligence and refinement in machine learning models. With this development, complementary investigation has been carried out to check if it is possible to beat the EMH and outperform the market using machine learning algorithms. As pointed out in [1] "…results have shown that machine learning methods could be successfully used toward stock predicting using stocks' historical data. Most of these existing approaches have focused on short term prediction using stocks' historical price and technical indicators". Also, it is known that investors use sentiment analysis to capture the market feelings and react quickly to them.

The vast majority of the work done in this machine learning field, use technical variables as inputs and very short operating periods, typically less than a day. One reason for this, is that technical analysis requires information that can be easily acquired and the possibility of automated trading systems, allows instant profits. The mayor disadvantage in this field, is that algorithmic trading is constantly evolving and, while doing day-trading, the algorithm proposed must beat others in order to obtain a profit (as it is supposed that price changes are non-significant during days). This means that a useful model can turn into obsolete in any time if not retrained. Critics often says that day-trading is just speculative noise to the market and does not provide any value for the real economy, as capital allocation usually goes away at the end of the day. This argument has been explained in [10], stating that AI race in the stock investing appears to be a zero-sum game for investors and the benefit of having a good model would be lost if that model is shared among other investors. Therefore, there are no incentives in sharing information. On the other hand, fundamental analysis of companies with machine learning has been much less used and investigated due to the difficulties in trustful data acquisition. However, focusing on those variables would allow for an objective capital allocation, in which collaboration between researchers and investors could be an incentive while also favouring the health of the economy.

An interesting methodology is explained in [1], where using only fundamental variables of companies, the selected portfolio outperforms the market using models such as Random Forest, Feed-forward Neural Network and Adaptive Neural Fuzzy Inference System. In that example, the predicting interval is one quarter and the available data are the companies in the S&P 100.

As the current situation, there are still many important issues to investigate in this field, including the aggregation of sentiment analysis with technical and fundamental analysis, addition of information about insiders' equity, macroeconomy variables and consumer's surveys.

4. DATA COLLECTION

As described in *section 3*, the most challenging part of generating a model that is useful to predict or select an optimal stock portfolio, is to collect data which is reliable and extensive enough. Two different types of variables have been gathered in this study.

4.1 Company's data

These variables are the main part of the study, in which medium/long-term tendencies will be captured by the model. For constructing a reliable data set, an Application Programming Interface (API[8]) key for the site *'End of day Historical Data'* (EOD) has been acquired.

EOD is a tech company that provides financial data API for analysists around the world, providing more than 150,000 tickers[9]. Some of the available APIs are listed below, see [11] for more information:

1. Fundamentals: access to fundamental data API for stocks, ETFs, Mutual Funds, and Indices from different exchanges and countries.
2. End of day: stock prices; daily, weekly and monthly data raw and adjusted to splits and dividends.
3. Real time: real-time data with a delay of less than 50ms via Web Sockets for the US market, FOREX, Cryptocurrencies.
4. Technical: technical variables for tickers on a specific period.
5. Economic: Government Bonds for more than 15 countries with different periods all over the world.

Fundamentals

The fundamentals API was used to retrieve all available annual data from tickers listed in the NYSE and the NASDAQ, obtaining almost 7,300 tickers from various sectors. The data was collected in *json* format.

For every ticker available in the API and listed in NYSE and NASDAQ, a *json* was retrieved with the following format (see *Annex A* for a summary example of ticker "AAPL"):

- General company information: sector, industry, exchange, etc.
- Date of report.
- Highlights: current key numbers to follow the stock.
- Valuation: book value and book ratios.
- Technical: current technical variables such as moving averages, beta ratio, short ratio, etc.
- Analysts' ratings: current ratings, target price and experts' expectations for the stock.
- Holders: information about institutional holding, insider holdings and insider transactions.

[8] Informatic bridge that enables different software systems to interact and share information.

[9] Unique series of letters and numbers used to identify a publicly traded company's stock or other financial instrument.

- Outstanding shares by date.
- Earnings: earnings made by the company presented annually or quarterly, with estimations of *eps* ratio and variations of that ratio over the periods.
- Financials: balance sheets and cash flow numbers over the history of the company. Variables are retrieved periodically in quarters or annual.

The transformation of the *json* file started by selecting the relevant features, for every company at each year in which the data is registered, general company data, outstanding shares, balance information and cash flow numbers are stored into a *pandas* data frame. The data set is constructed in a way that an instance is a ticker in a given year (as the data is retrieved annually). It is important to notice that each company has a fixed day of the year in which the report, balance sheet and cash flow statements are made public. That date is collected as well to join with other data later.

Price

Once obtained the company's fundamentals, the prices are key to construct the data set. Using the *End of Day* API, prices for every day can be obtained. As explained in [11], prices obtained can be adjusted by splits and dividends for a correct analysis of the price fluctuation over the years. The variables retrieved daily for a selected period are:

- Date.
- Open: opening price.
- High: highest price reached in the selected date.
- Low: lowest price reached in the selected date.
- Close: closing price.
- Adjusted close: closing price adjusted by splits and dividends.
- Volume: number of stocks traded in that day.

At an instance created before with fundamental variables, a monthly average of the closing price and adjusted closing price are joined into the data set. A monthly average of the prices is selected instead of a single closing price to avoid peaks in the prices, which is usual to happen in days where reports are issued.

Fundamental ratios construction

With the balance sheet and cash flow statements, as well as the prices for those dates, the ratios explained in *Section 2.1* can be constructed. Note that some simplifications have been done when calculating EPS, ROE and QR.

$$eps = \frac{netIncome}{commonStockSharesOutstanding} \tag{4-1}$$

$$per = \frac{price * commonStockSharesOutstanding}{netIncome} \tag{4-2}$$

$$der = \frac{totalCurrentLiabilities}{totalStockholderEquity} \tag{4-3}$$

$$roe = \frac{netIncome}{totalStockholderEquity} \tag{4-4}$$

$$wcr = \frac{totalCurrentAssets}{totalCurrentLiabilities} \tag{4-5}$$

$$qr = \frac{totalCurrentAssets - inventory}{totalCurrentLiabilities} \tag{4-6}$$

$$fcfps = \frac{freeCashFlow}{commonStockSharesOutstanding} \tag{4-7}$$

$$pcfr = \frac{fcfps}{price} \tag{4-8}$$

Also, for all the ratios proposed, the differences (first and second differences) have been calculated:

$$ratio\_diff1_j = ratio_j - ratio_{j-1} \tag{4-9}$$

$$ratio\_diff2_j = ratio_{j-1} - ratio_{j-2} \tag{4-10}$$

4.2 Macroeconomy data

The second part of the dataset is to collect macroeconomic variables. The reason behind the acquisition of those variables, is to study if ML models are able to predict global market tendencies in bearish or bullish years.

This data collection was made though the Federal Reserve Economic Data (FRED) API, which has access to an economic database maintained by the Federal Reserve Bank of St. Louis. FRED provides access to a collection of economic and financial data for the US including economic indicators such as interest rates, GDP, employment, inflation, etc.

The variables that were acquired are the following (see [12] for more information):

1. Population: represents the estimated total resident population of the US and its territories [13].
2. GDP: the real gross domestic product is the inflation adjusted value of the goods and services produced by labour and property located in the United States (Billions of dollars) [14].
3. Inflation rate: percentage change in the prices of goods and services in the United States economy over time [15].
4. Money supply: represents the seasonally adjusted measure of M2 money stock, which includes M1 (currency, demand deposits, and other checkable deposits) plus savings deposits, small-denomination time deposits, and retail money market mutual fund shares (Millions of US dollars) [16].
5. Employment: refers to the seasonally adjusted total number of employees on nonfarm payrolls, which includes workers in industries such as manufacturing,

construction, retail, healthcare, and other services, but excludes workers in farming, private households, and non-profit organizations (Thousands of jobs) [17].

6. Unemployment rate: represents the number of unemployed as a percentage of the labour force [18].
7. Interest rates: two different time series for interest rates, the "3-Month Treasury Bill: Secondary Market Rate" and "10-Year Treasury Constant Maturity Rate". The first one, is the interest rate for new 3-month Treasury bills that are sold on the secondary market. The latter is the interest rate for a 10-year Treasury bond with a constant maturity [19], [20].

All the time series were retrieved annually and the first difference for all the macroeconomic variables are calculated. A *pandas* data frame was generated from the data, in which the years were the instances and columns were the variables described before.

4.3 Constructing the data set

After having the fundamentals of the companies on one data frame and the macroeconomy variables of all years on the other, the construction of the data set was made. For each instance, that is characterized by a tuple of *(date, ticker)* and have all the fundamentals; the macroeconomic variables are added. Note that reports can be issued at different dates for different companies, but macroeconomic variables are gathered by the end of the year. The addition of the macroeconomic variables is done in such a way that the values are linearly interpolated to be accurate for the time of the year in which the report was issued.

Is key to note that macroeconomic variables are introduced in the data set considering that the time dependence is not violated, meaning that for a given date, the macroeconomic values are calculated for the period of one year before that.

Finally, the “target” variable is introduced into the data set as the change in price for the same ticker at the following year, see *Equation (4-11).*

$$target_j = \frac{price\_adjisted_{j+1} - price\_adjusted_j}{price\_adjusted_j} * 100 \qquad \textit{(4-11)}$$

The obtained data set then has 65,584 instances and 73 features which are described in *ANNEX B.*

5. ANALYSIS

This section explains the procedure and reasons that were followed to generate a ML model that can be used in stock investment. For generating a ML model, the following steps were followed: data preprocessing, ML models approaches, model selection and hyperparameter tuning for the final model.

5.1 Data preprocessing and model evaluation

In this section, the data is prepared to be fed to a ML model and then split in order to generate the correct workflow for evaluating the results.

Data set filtering

First, it is important to filter the data set from instances that can induce noise in the model and clear features and instances with excessive missing data. The following steps were followed for filtering the data set:

1. Remove all instances with missing *target.* The *target* is essential to construct a reliable data set and it cannot be accurately imputed, specifically in this case where the target present an excessive deviation with high skewness.
2. Remove all instances of companies which are not included in S&P 500 index.
3. Remove instances where:

   $$target > 10{,}000 \qquad \textit{(5-1)}$$
4. Replace the target variable where:

   $$if: 1{,}000 < target \leq 10{,}000 \rightarrow target = 1{,}000 \qquad \textit{(5-2)}$$
5. Remove instances with more than 40% of features missing.
6. Remove feature with more than 40% of the instances missing.

Those important steps in filtering the data have two main objectives, the first one is to remove features and instances in which missing data are excessive. The reason for deleting those data is because excessive missing data might end up in poor imputing in the second stage of this preprocessing, leading to noisy instances. The second objective is not straight forward and has been done with iterative modelling. As it has been explained before, one of the main problems of generating useful ML models for stocks, is the extremely high variability in stock markets, which sometimes have nothing to relate with fundamental variables of companies. It could be the news, a review or an international conflict that can make stock prices go up and down in seconds. To reduce this effect, selecting only high market capitalization companies is a good strategy and truncating the *target* can be used as well.

The obtained filtered data set has a total amount of 9,784 instances and 70 features.

Categorical encoding, scaling and imputation

For the years 2000 to 2021, the expanding window method is performed generating 22 different pairs of training and testing sets. The expanding window is set such that for the year *i* being generated, the ML model is trained with instances with years ≤ *i* and tested with instances coinciding with year *i+1.*

For every one of the 22 pairs of train and test data sets that are being generated, the following pipeline[10] is applied to the data (algorithms used can be found on [21]).

1. One-hot encoder [22]: this encoder is only used in the variable *Sector*.
2. Target encoder [23]: for every other categorical variable in the data set (*Ticker*, *Industry*).
3. Standard scaler [24]: using mean and standard deviation as the scaler parameters.
4. KNN imputer [25]: using five neighbours.

In the figure below, it can be seen the heat map of the correlation matrix for the whole dataset after preprocessing (for year 2021). Note: target encoding has been used for feature *Sector* in this case as well, to ease the visualization of the matrix, see features represented in *ANNEX C*.

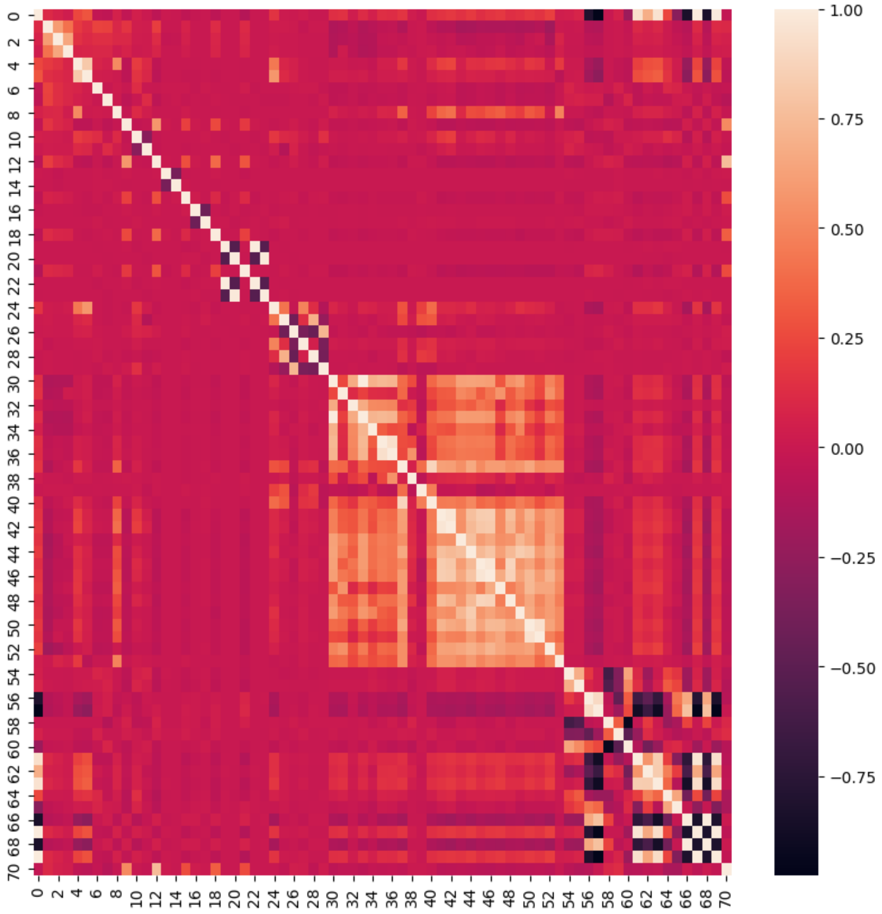

Fig. 5-1: Correlation matrix for training set year 2022.

5.2 Different approaches

In this section, some of the approaches used for generating the most useful ML model are shown and explained. Also, the advantages and disadvantages are explained in detail.

[10] Sequence of data processing steps that are chained together.

5.2.1 Prediction model

The initial objective of this thesis and first model that was to predict stock prices, in which the target variable is predicted by a regressor model. In this model, the *target* variable was transformed in order to avoid skewness and high variability.

$$target_{transformed} = \ln\left(\frac{target}{100} + 1\right) \quad (5\text{-}3)$$

After applying the transformation, four different regression models were fitted to the 22 train sets available and after, tested in the corresponding test sets. Four models were tried (see [21] for general information):

1. Random forest [26]: with *min_samples_split = 0.025* and default values for the other hyperparameters.
2. Extra-trees [27]: with *min_samples_split = 0.025* and default values for the other hyperparameters.
3. Gradient boosting [28]: with *min_samples_split = 0.025* and default values for the other hyperparameters.
4. Support vector regressor [29]: default hyperparameters.

However, the prediction model as a regression of the transformed target variable did not return useful results. When comparing the four models with respect to a naïve regressor [30], we can see the following errors in the 22 different models that have been generated (see *Table 5-1*).

Model	Mean squared error
Naïve regressor	0.120±0.134
Random forest	0.158±0.154
Extra-trees	0.149±0.151
Gradient boosting	0.148±0.148
Support vector regressor	0.130±0.134

Table 5-1: Regression models´ errors.

The models do not behave well, giving similar errors than the naïve regressor, with high variation from year to year. However, it could be said that when investing, it is not necessary to predict exactly how much will be the price of a company the following year, but to select some companies to build a strong portfolio which beats the market and obtain the desired results, which can vary depending on the investor. A similar idea has been used in [1].

Following that idea, we could select the best 25 companies (highest predicted values) each year for each model and observe how those companies behave over time. In the *Fig. 5-2*, compound return-initial investment ratio are shown for the different models, the average of the whole data set and the S&P 500 index.

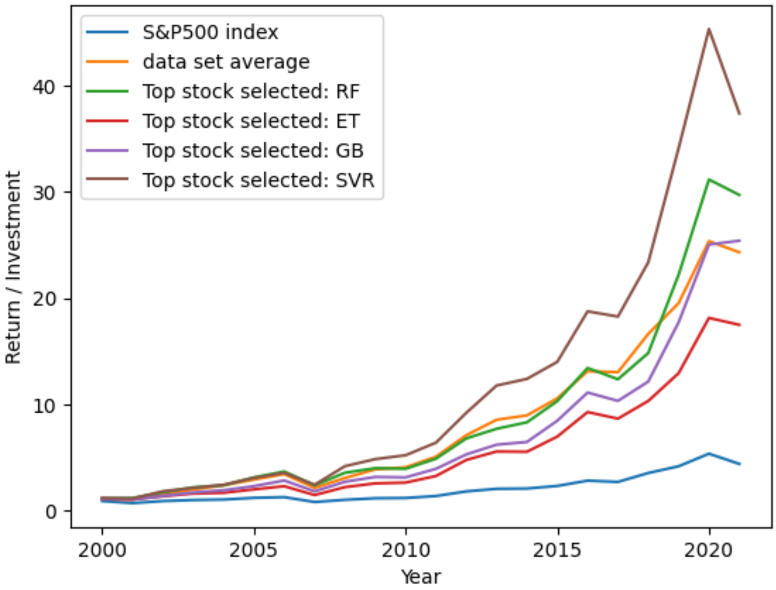

Fig. 5-2: Accumulated returns over the years for top 25 companies.

Surprisingly, the S&P 500 compound gives lower returns than the average of the companies in the data set, however, the index shows high stability in recession years such as 2007 (predicting 2008 crisis). Other insight that can be obtained, is that the SVR model achieves better performance than the average of the market, returning a 20.24% average annually. The other models seem to achieve similar results than the average of the data, with 17.34% average return.

A hypothesis test can be used to prove the dominance of the SVR over the data set's average. Considering that the difference between returns follows a Gaussian distribution, we can use a Student's *t-test* to check:

$\mathrm{X} = \text{"return of the SVR model"};\ \mathrm{Y} = \text{"return of the average test set"}$

$\mathrm{H0)}\ \mathrm{E(X-Y)}\ =\ 0$

$\mathrm{H1)}\ \mathrm{E(X-Y)} \geq 0$

After computing the *t-test* results show that is not possible to discard the null hypothesis, as the $t_{22-1} = 1.29$ resulting $\alpha = 0.11$.

5.2.2 Classification model

The second strategy for generating useful models that can bring insights about stock markets, is to fit a classification model over the data set. For doing that, the target variable was transformed into a binary variable following:

$$target_{transformed} = \begin{cases} 1\ if\ target \geq 0 \\ \ 0\ otherwise \end{cases} \tag{5-4}$$

The idea behind generating a classification model instead of a regressor one, is to simplify the model and the information that the model is fed with. As discussed before, the target variable is highly unstable, with typical values that are in the range of -10 to 20, but in some cases, it can take values greater than 1,000. Also, those unexpected high return, could add noise to the model which tries to fit those unrealistic high values, losing interest in capturing the trends of the typical cases.

Four different models were applied to the classification approach, following the same idea of the expanding window and the 22 different years. The models were the following (see [21] for general information):

1. Extra-trees [31]: with *min_samples_split = 0.025* and default values for the other hyperparameters.
2. K-nearest neighbours [32]: with *n_neighbors = 30* and default values for the other hyperparameters.
3. Support vector classifier [33]: with default hyperparameters and the *probability = True* tool.

Note that the three models that were fitted, are able to return probabilities for each class. Also, we can compare the accuracy of the three models with respect to a naïve classifier [34]. In *Table 5-2* an average (and standard deviation) of the accuracies can be found for different models.

Model	**Accuracy**
Naïve	0.709±0.229
Extra-trees	0.688±0.216
K-nearest neighbours	0.684±0.213
Support vector classifier	0.673±0.201

Table 5-2: Classification models' accuracies.

Although accuracies obtained are around 0.70, none of the models are able to outperform the naïve classifier, which is set to predict for all the instances, the majority class. Note that in this case the majority class is *1*, meaning that in the whole data set, around 70% of the companies increases its stock price.

But as the models give probabilities also, we could use a similar idea as it was proposed in *section 5.2.1* and select 25 companies which are the most probable companies of increasing its value. After that selection, we can calculate an average return for each model and compound it over the years. In *Fig. 5-3*, it can be seen the returns-initial investment ratio for the different models, the average of the whole data set and the S&P 500 index.

Those results show how every model that was fitted is consistently over the average of the data set and the S&P 500 compound index as well. Also, we could say that Extra-trees, that is the best model, obtained an average return of 20.69% in contrast with the average of the market which is 17.34%.

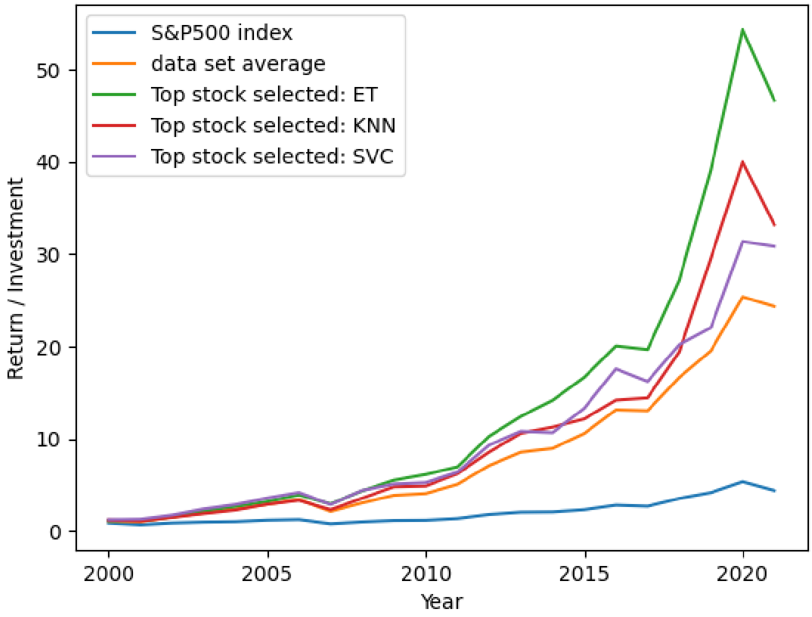

Fig. 5-3: Accumulated returns over the years for top 25 companies.

For the returns achieved by the Extra-trees, a Student's *t-test* was calculated in order to check if the model is able to consistently beat the market and overrule the EMH.

$\mathrm{X} = \text{"return of the ET model"};\ \mathrm{Y} = \text{"return of the average test set"}$

$\mathrm{H0})\ \mathrm{E}(\mathrm{X}-\mathrm{Y}) = 0$

$\mathrm{H1})\ \mathrm{E}(\mathrm{X}-\mathrm{Y}) \geq 0$

After computing the *t-test* results show that $t_{22-1} = 1.70$ resulting $\alpha = 0.05$. Assuming $\alpha^* = 0.05$, it is concluded that there is not statistical significance for discarding the null hypothesis. However, it can be noted that a slight optimization of the model will lead into statistical significance.

5.3 Definitive approach and evaluation

After computing the different types of models, the Extra-trees is the model that reaches the highest average return and compound return over the years.

The definitive approach contemplates the hyperparameter tuning of the Extra-trees classifier model to calculate an approximate return of the model being trained. The hyperparameters to be optimized are:

- Number of decision trees in the ensemble is set in the model as *n_estimators* [31] and is a direct hyperparameter of the Extra-trees model that will be tuned for the whole set of the 22 models created, one for each year. This hyperparameter governs the randomness of the model.
- The minimum number of samples required to split an internal node, governs the complexity of each tree in the ensemble and is set with the parameter *min_samples_split* [31]. It is also a direct hyperparameter of the Extra-trees classifier.

- Number of selected companies in the portfolio is the most important hyperparameter of the model and it is a trade-off between risk and reward. For small values of selected companies, the reward tends to be greater but more unstable. The key is to find the optimum number of companies which offers the greater return in the long term.

For generating this hyperparameter tuning, the test set of each year will be randomly divided into validation and test at a 50/50 proportion. The hyperparameters have been selected in order to maximize the average return of the validation set using a grid search with the following values:

- Number of decision trees: [50, 100, 200, 500, 1000, 2000]
- Minimum number required for splitting a node: [0.001, 0.005, 0.01, 0.05]
- Number of selected companies: [10, 15, 20, 25]

The grid is proposed taking into consideration the model characteristics, time consumption and investments constrains. For the number of companies, 10 is considered to be the lowest value possible in order to avoid variability issues (see *section 5.4*) and excessive risks. On the other hand, 25 companies are the maximum set for the portfolio, which is thought for a small investor's possibilities.

A summary of the best results achieved with the grid search are showed in *Table 5-3.* Note that the variability of these kind of models is not taken into consideration in this study because of the expensive calculations involved but is addressed in *section 5.4.*

n_estimators	nin_samples_split	N° of comp.	Avg. ret. val. (%)	Avg. ret. test (%)
1000	0.001	10	24.31	23.45
100	0.001	20	22.89	19.75
1000	0.001	15	22.83	21.99

Table 5-3: Grid search top 3 results.

As it can be seen in the results, large number of trees generate better results, the opposite of what happens with the number of samples required to split a node, in which lower values optimize the model. The number of companies that optimizes the validation metric is 10, which is a sensible number for a small investor's portfolio. Lower numbers of selected companies have not been considered for being too risky and volatile.

Again, a Student's *t-test* can be calculated to check if the model is able to consistently beat the market and overrule the EMH. This calculation was computed over the returns on the test set.

$X = \text{"return of the ET, tuned model"};\ Y = \text{"return of the average test set"}$

$H0)\ E(X - Y) = 0$

$H1)\ E(X - Y) \geq 0$

After computing the *t-test* results show that $t_{22-1} = 1.76$ resulting $\alpha = 0.046$. Assuming $\alpha^* = 0.05$, it is concluded that there is statistical significance for discarding the null hypothesis and assume that the model outperforms the average of the market.

5.4 Final model

The last step in the ML workflow, is to generate a model that can be used to select stocks for the future, in this case, the hyperparameter tuning can be done in order to optimize the metric for the whole test set, without the necessity of splitting into validation and test. A similar procedure as the explained in *section 5.3,* has been followed for hyperparameters tuning with grid search.

For this case, it has been noticed that the models generated are subject to sensitive variations in the obtained metrics when changing intrinsic value as the seed[11]. This phenomenon is due to the nature of the model being generated, where the ensemble model predicts probabilities, and then, over a selection of the instances, the returns are calculated. It is the case that a slight change in a probability, end up in including or excluding companies, which changes the result of the investment. The model ends up being slightly unstable, with small changes in the input generating considerable changes in the output.

For measuring that variability generated by the model, five different models have been trained with 5 different seeds over a grid search. In this case, not only the average of the metrics is going to be maximized, but also a consideration of the variability given by those hyperparameters. The metric to maximize is then: $\overline{return} - \sigma_{return}$ (mean – standard deviation). The top three results are shown in *Table 5-4.*

n_estimators	nin_samples_split	N° of companies	Avg. ret. (%)	σ_{return}	metric
2000	0.001	25	22.41	0.47	21.94
2000	0.005	25	21.66	0.42	21.24
200	0.005	20	22.31	1.16	21.14

Table 5-4: Grid search results for prediction model.

The obtained model then, contemplates the investment strategy of maximizing the results, but with a model that is considered with low variability. From the results, we can conclude that models with higher values of selected companies and numbers of trees in the ensemble, favours the reduction in the deviation of the results. A complete version of *Table 5-4*, can be found on *ANNEX D.*

For the final model with tuned hyperparameters, companies that were selected though the years are shown in *ANNEX E* and the compound returns are shown in *Fig. 5-4.*

[11] Parameter of the model that is used for reproducibility when generating random numbers.

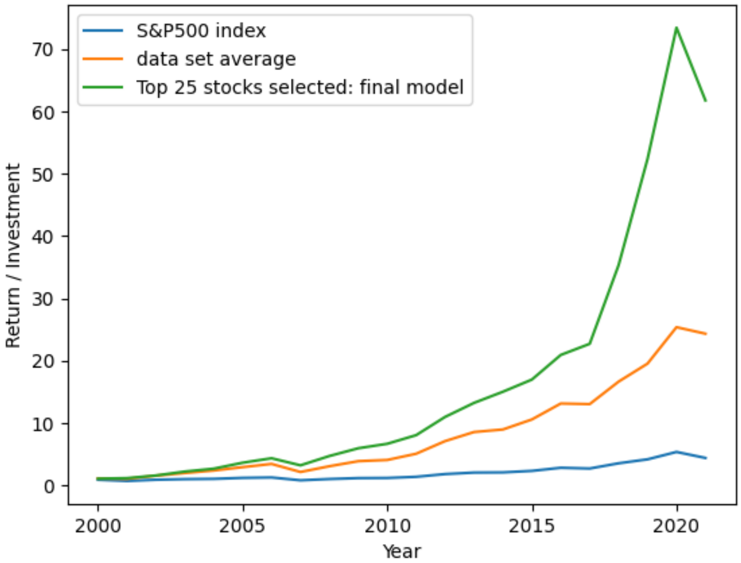

Fig. 5-4: Accumulated returns over the years for top 25 companies, final model.

6. CONCLUSIONS

6.1 Results and discussions

In this study, machine learning has been used to generate models for stock investing using fundamental variables of companies and macroeconomic data. Those models have been used to generate insights over the different approaches that an investor can use on the medium/long term.

Three different approaches have been used: stock price prediction, stock price classification and stock selection. Expanding window from 2000 to 2021, with 1 year time period, has been used in all cases for avoiding future data and target leakage.

When generating stock prediction, the objective is to generate a regression model that can predict the future price of companies in a 1-year period. Generated ML models MSE are between 8.33 to 31.60 % above the errors obtained with a naive regressor, concluding that none of the models are able to predict the target variable.

Classification models are fitted to predict if a given company will increase its stock price in the next year, but as in the previous case, none of the models were able to generate better metrics than the naive. ML models' accuracies were 3.60 to 2.10 % less than the naive classifier.

However, when using the previous models to select stocks, it is shown that better returns than the data set's average and the S&P 500 compound, are obtained. This approach works by selecting the best companies in the classification/regression models, to obtain a portfolio and observe how that portfolio performs over 1 year period, over the 22 years included in the expanding window.

The definitive model then, is to fit a ET classification model and use it to portfolio selection. The return of those companies is computed for every year in the expanding window and then compared with the average of all the companies. A grid search hyperparameter tuning has been carried out, optimizing the parameters: number of trees in the ensemble, minimum number of instances required to split a node and number of companies selected in the portfolio. For computing the grid search, the future data in the expanding window has been split into test and validation sets. The grid search is set in order to maximize the validation average return. The obtained tuned model reaches a 23.45% of average return in the test set, over the period 2000-2022, in comparison with 17.34% average return of the whole data set. Moreover, a t-test has been computed and concluded that there is statistical significance to conclude that the model outperforms the market consistently and therefore, discard the EMH.

Finally, the final model has been computed considering the average return but also the variability of the models. In that study, it was found out that small number of selected companies and small number of trees, tend to generate more unstable models, with small changes in the inputs, generating high impact on the returns. The optimized model seeks a trade-off between performance and stability, obtaining 22.41% of average return and a standard deviation of 0.47%.

6.2 Further lines of work

As vast as it can be, this subject offers different ways of approaching the stock investment strategies with ML. Undoubtedly, fundamental variables of companies are the core of the data used in this study. Some further work pointing in this direction, could include:

1. Different time prediction horizon can be analysed to optimize the hyperparameter of the model that is fixed to 1 year in this study. If using quarterly data, changing the time horizon could be possible and because of this, more vast data sets are achieved. In this sense, when using a larger prediction horizon, more accurate results with low variability should be obtained, diluting the peaks that induce noise in the models. On the other hand, if the horizon is too large, initial data might end up being obsolete for the end of the period.
2. Technical variables can be added to the model to search for low investing entering prices and tendencies. It would bring important insights to compute a model with both fundamental and technical variables and compare the importance of the variables used in the analysis. As discussed before, technical analysis has been widely studied, but the research for ML in stock context for medium/long term is yet to be perfected.
3. A explained in the study, the construction of the data set is key to obtain useful models. New ratios can be created from the fundamental variables to seek a better performance. Aligned with thought, it is considered important to generate a feature selection of the data set and generate a sense of the importance of the features.

BIBLIOGRAPHY

[1] Huang, Y., Capretz, L. F., & Ho, D. (2021). Machine Learning for Stock Prediction Based on Fundamental Analysis. Retrieved from: https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1569&context=electricalpub

[2] How Do Companies Get Listed on the New York Stock Exchange? Bob Haring (2009), [Online]. Available: https://finance.zacks.com/companies-listed-new-york-stock-exchange-7015.html

[3] World Bank. (2023). Market capitalization of listed companies (current US$). Retrieved from: https://data.worldbank.org/indicator/CM.MKT.LCAP.CD

[4] Statista. (2023). Global stock markets by country. Retrieved from: https://www.statista.com/statistics/710680/global-stock-markets-by-country/

[5] Lynch, P. (1989). One Up on Wall Street: How to Use What You Already Know to Make Money in the Market. Simon & Schuster.

[6] Malkiel, B. G. (2003). The Efficient Market Hypothesis and Its Critics. Princeton University Press. Retrieved from: https://eml.berkeley.edu/~craine/EconH195/Fall_16/webpage/Malkiel_Efficient%20Mkts.pdf

[7] Maverick, J.B. (2022). The Weak, Strong, and Semi-Strong Efficient Market Hypotheses. Investopedia. Retrieved from: https://www.investopedia.com/ask/answers/032615/what-are-differences-between-weak-strong-and-semistrong-versions-efficient-market-hypothesis.asp

[8] Drutarovská, J. (2014). Speculative Activities in the Financial Markets and Its Relation to the Real Economy. Retrieved from: https://www.jopafl.com/uploads/issue6/SPECULATIVE_ACTIVITIES_IN_THE_FINANCIAL_MARKETS_AND_ITS_RELATION_TO_THE_REAL_ECONOMY.pdf

[9] Cattlin, R. (2021). What are Economic Indicators. City Index. Retrieved from: https://www.cityindex.com/en-sg/news-and-analysis/what-are-economic-indicators/

[10] Strader, T. J. (2020). Machine Learning Stock Market Prediction Studies: Review and Research Directions. Retrieved from: https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=1435&context=jitim

[11] End of day historical data: https://eodhistoricaldata.com/financial-apis/

[12] Federal Reserve Economic Data (FRED) API: https://fred.stlouisfed.org/docs/api/fred/

[13] U.S. Census Bureau, Total Population: All Ages including Armed Forces Overseas [POP], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/POP

[14] U.S. Bureau of Economic Analysis, Gross Domestic Product [GDP], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GDP

[15] Organization for Economic Co-operation and Development, Consumer Price Index: Total All Items for the United States [CPALTT01USM657N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPALTT01USM657N

[16] Board of Governors of the Federal Reserve System (US), M2 [M2SL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/M2SL

[17] U.S. Bureau of Labor Statistics, All Employees, Total Nonfarm [PAYEMS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PAYEMS

[18] U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE

[19] Board of Governors of the Federal Reserve System (US), 3-Month Treasury Bill Secondary Market Rate, Discount Basis [TB3MS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/TB3MS

[20] Board of Governors of the Federal Reserve System (US), Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity, Quoted on an Investment Basis [GS10], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/GS10

[21] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[22] Repository: scikit-learn/scikit-learn. (2023). preprocessing/_encoders.py. Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/preprocessing/_encoders.py#L215

[23] Module: category_encoders.target_encoder. (n.d.). Retrieved from: https://contrib.scikit-learn.org/category_encoders/_modules/category_encoders/target_encoder.html#TargetEncoder

[24] Module: sklearn.preprocessing._data. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/preprocessing/_data.py#L644

[25] Module: sklearn.impute._knn. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/impute/_knn.py#L20

[26] Module: sklearn.ensemble._forest. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/ensemble/_forest.py#L1441

[27] Module: sklearn.ensemble._forest. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/ensemble/_forest.py#L2132

[28] Module: sklearn.ensemble._gb. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/ensemble/_gb.py#L1418

[29] Module: sklearn.svm._classes. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/svm/_classes.py#L1104

[30] Module: sklearn.dummy. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/dummy.py#L445

[31] Module: sklearn.ensemble._forest. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/ensemble/_forest.py#L1778

[32] Module: sklearn.neighbors._classification. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/neighbors/_classification.py#L24

[33] Module: sklearn.svm._classes. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/svm/_classes.py#L554

[34] Module: sklearn.dummy. (n.d.). Retrieved from: https://github.com/scikit-learn/scikit-learn/blob/364c77e04/sklearn/dummy.py#L25

[35] Francisco Esteves Muñoz (2023). masters_thesis_stock_selection/code (Version 1) Retrieved from: https://github.com/festevesmunoz/masters_thesis_stock_selection

ANNEX A

A summary of the json file retrieved by the fundamentals API for the company ‘Apple’ (ticker AAPL).

{
"General": {"Code": "AAPL”, “Type": "Common Stock", "Name": "Apple Inc", "Exchange": "NASDAQ", "CurrencyCode": "USD", "CurrencyName": "US Dollar", "CurrencySymbol": "$", "CountryName": "USA", "CountryISO": "US", … , "Description": "...", "Address": "...", "AddressData": {...}, "Listings": {...} },
"Officers": {...},
… ,
"Highlights": {"MarketCapitalization": 2413630849024, "MarketCapitalizationMln": 2413630.849, "EBITDA": 125287997440, "PERatio": 25.8998, "PEGRatio": 2.7503, "WallStreetTargetPrice": 169.39, "BookValue": 3.581, "DividendShare": 0.91, "DividendYield": 0.006, "EarningsShare": 5.89, "EPSEstimateCurrentYear": 5.97, "EPSEstimateNextYear": 6.58, "EPSEstimateNextQuarter": 1.48, "EPSEstimateCurrentQuarter": 1.94, …},
"Valuation": {"TrailingPE": 25.8998, "ForwardPE": 23.0947, "PriceSalesTTM": 5.5067, "PriceBookMRQ": 44.6301, "EnterpriseValue": 2333263054880, "EnterpriseValueRevenue": 5.9171, "EnterpriseValueEbitda": 17.5251},
"SharesStats": {"SharesOutstanding": 15821899776,"SharesFloat": 15805016518,"PercentInsiders": 0.074, "PercentInstitutions": 61.356, "SharesShort": null, "SharesShortPriorMonth": null, "ShortRatio": null, “ShortPercentOutstanding": null, "ShortPercentFloat": null},
"Technicals": {"Beta": 1.2779, "52WeekHigh": 178.5285, "52WeekLow": 123.9807, "50DayMA": 140.098, "200DayMA": 147.3938, "SharesShort": 115480341, "SharesShortPriorMonth": 121868919, "ShortRatio": 1.6, "ShortPercent": 0.0073},
"SplitsDividends": {…},
"AnalystRatings": {…},
"Holders": {"Institutions": {...}, "Funds": {...}},
"InsiderTransactions": {...},
"outstandingShares": {"annual": {"0": {"date": "2022", "dateFormatted": "2022-12-31","sharesMln": "15955.7180", “shares": 15955718000}, "1": {"date": "2021", "dateFormatted": "2021-12-31","sharesMln": "16519.2910", "shares": 16519291000.000002}, ...},
"quarterly": {...}},
"Earnings": {"History": {"2022-12-31": { "reportDate": "2023-02-02", "date": "2022-12-31", "beforeAfterMarket": "AfterMarket", "currency": "USD", "epsActual": 1.88, "epsEstimate": 1.94, "epsDifference": -0.06, "surprisePercent": -3.0928 }, "2022-09-30": {"reportDate": "2022-10-27", "date": "2022-09-30", "beforeAfterMarket": "AfterMarket", "currency": "USD", "epsActual": 1.29, "epsEstimate": 1.27, "epsDifference": 0.02, "surprisePercent": 1.5748 }, ...}, "Trend": {...}, "Annual": {...} },
"Financials": {
"Balance_Sheet": {
"currency_symbol": "USD",
"quarterly": {...},
"yearly": {
"2022-09-30": { "date": "2022-09-30", "filing_date": "2022-10-27", "currency_symbol": "USD", "totalAssets": "352755000000.00", "intangibleAssets": null, "earningAssets": null, "otherCurrentAssets": "21223000000.00", "totalLiab": "302083000000.00", "totalStockholderEquity": "50672000000.00", "deferredLongTermLiab": null, "otherCurrentLiab": "60845000000.00", "commonStock": "64849000000.00", "capitalStock": "64849000000.00", "retainedEarnings": "-3068000000.00", "otherLiab": "49142000000.00", "goodWill": null, "otherAssets": "54428000000.00", "cash": "23646000000.00", "cashAndEquivalents": "5100000000.00", "totalCurrentLiabilities": "153982000000.00", "currentDeferredRevenue": "7912000000.00", "netDebt": "96423000000.00", "shortTermDebt": "21110000000.00", "shortLongTermDebt": "21110000000.00",

```
"shortLongTermDebtTotal": "120069000000.00",  "otherStockholderEquity": "-11109000000.00",
"propertyPlantEquipment": "42117000000.00",  "totalCurrentAssets": "135405000000.00",
"longTermInvestments": "120805000000.00",  "netTangibleAssets": "50672000000.00",
"shortTermInvestments": "24658000000.00",  "netReceivables": "60932000000.00", "longTermDebt":
"98959000000.00",  "inventory": "4946000000.00", "accountsPayable": "64115000000.00",
"totalPermanentEquity": null, "noncontrollingInterestInConsolidatedEntity": null,
"temporaryEquityRedeemableNoncontrollingInterests": null, "accumulatedOtherComprehensiveIncome": "-
11109000000.00",  "additionalPaidInCapital": null,  "commonStockTotalEquity": "64849000000.00",
"preferredStockTotalEquity": null,  "retainedEarningsTotalEquity": "-3068000000.00",  "treasuryStock": null,
"accumulatedAmortization": null,  "nonCurrrentAssetsOther": "54428000000.00",
"deferredLongTermAssetCharges": null,  "nonCurrentAssetsTotal": "217350000000.00",
"capitalLeaseObligations": null, "longTermDebtTotal": "99771000000.00", "nonCurrentLiabilitiesOther":
"49142000000.00",  "nonCurrentLiabilitiesTotal": "148101000000.00”, “negativeGoodwill": null, “warrants": null,
"preferredStockRedeemable": null, “capitalSurpluse": "64848840570.00",  "liabilitiesAndStockholdersEquity":
"352755000000.00",  "cashAndShortTermInvestments": "48304000000.00",
"propertyPlantAndEquipmentGross": "114457000000.00", "propertyPlantAndEquipmentNet": "42117000000.00",
"accumulatedDepreciation": "72340000000.00",  "netWorkingCapital": "-18577000000.00",
"netInvestedCapital": "170741000000.00",  "commonStockSharesOutstanding": "16325819000.00"},
          ...
        },
        "Cash_Flow": {"currency_symbol": "USD”, “quarterly": {...},
            "yearly": { "2022-09-30": { "date": "2022-09-30",  "filing_date": "2022-10-27", "currency_symbol":
USD",  "researchDevelopment": "26251000000.00", "effectOfAccountingCharges": null, "incomeBeforeTax":
"119103000000.00", "minorityInterest": null, "netIncome": "99803000000.00",
"sellingGeneralAdministrative": "25094000000.00",  "sellingAndMarketingExpenses": null, "grossProfit":
"170782000000.00",  "reconciledDepreciation": "11104000000.00", "ebit": "122034000000.00", "ebitda":
"133138000000.00", "depreciationAndAmortization": "11104000000.00", "nonOperatingIncomeNetOther": "-
334000000.00", "operatingIncome": "119437000000.00", "otherOperatingExpenses": "274891000000.00",
"interestExpense": "2931000000.00", "taxProvision": "19300000000.00", "interestIncome": "106000000.00",
"netInterestIncome": "-106000000.00",  "extraordinaryItems": null,  "nonRecurring": null, "otherItems": null,
"incomeTaxExpense": "19300000000.00",  "totalRevenue": "394328000000.00", "totalOperatingExpenses":
"51345000000.00", "costOfRevenue": "223546000000.00",  "totalOtherIncomeExpenseNet": "-334000000.00",
"discontinuedOperations": null,  "netIncomeFromContinuingOps": "99803000000.00",
"netIncomeApplicableToCommonShares": "99803000000.00", "preferredStockAndOtherAdjustments": null
          },
          ...
        }}}}
```

ANNEX B

The obtained data set after performing the data collection, result in the following variables:

- year: year in which the report was issued.
- Ticker: code for identifying the company.
- Industry: industry of the company.
- Sector: sect
- Price: unadjusted price averaged in the month of the report submission (USD).
- price_adjusted: adjusted price by splits and dividends averaged in the month of the report submission (USD).
- price_diff1: current difference of the adjusted price (% of change).
- price_diff2: previous fiscal year difference of the adjusted price (% of change).
- market_cap: market capitalization of the company (USD billions).
- eps: earnings per share ratio.
- eps_diff1: earnings per share ratio current variation.
- eps_dif2: earnings per share ratio previous variation.
- per: price-earnings ratio.
- per_diff1: price-earnings ratio current variation.
- per_diff2: price-earnings ratio previous variation.
- der: debt-earnings ratio.
- der_diff1: debt-earnings ratio current variation.
- der_diff2: debt-earnings ratio previous variation.
- wcr: working capital ratio.
- wcr_diff1: working capital ratio current variation.
- wcr_diff2: working capital ratio previous variation.
- qr: quick ratio.
- qr_diff1: quick ratio current variation.
- qr_diff2: quick ratio previous variation.
- fcfps: free cash flow per share.
- fcfps_diff1: free cash flow per share current variation.
- fcfps_diff2: free cash flow per share previous variation.
- pcfr: price-cash flow ratio.
- pcfr_diff1: price-cash flow ratio current variation.
- pcfr_diff2: price-cash flow ratio previous variation.
- totalAssets (USD).
- IntangibleAssets (USD).
- otherCurrentAssets (USD).
- totalLiab: total liabilities (USD).
- otherCurrentLiab (USD).

- beginPeriodCashFlow: cash flow at the beginning of the period (one fiscal year before) (USD).
- endPeriodCashFlow: cash flow at the report issue date (USD).
- totalCashFromOperatingActivities: cash flow from operating activities (USD).
- changeInCash: difference between current cash flow and previous fiscal year (USD).
- otherCashflowsFromInvestingActivities (USD).
- changeInWorkingCapital: change in working capital for that period (USD).
- freeCashFlow: current free cash flow (USD).
- researchDevelopment: investment on research and development (USD).
- incomeBeforeTax: income before taxes generated on the last fiscal year (USD).
- netIncome: earnings generated in the last fiscal year (USD).
- sellingGeneralAdministrative (USD).
- grossProfit: gross income generated in the last fiscal year (USD).
- ebit: earnings before interest and taxes in the last fiscal year (USD).
- ebitda: earnings before interest, taxes, depreciation and amortization in the last fiscal year (USD).
- depreciationAndAmortization: depreciation and amortization over the last fiscal year (USD).
- operatingIncome (USD)
- totalOperatingExpenses (USD)
- taxProvision: expected taxes to be paid in the current fiscal year (USD).
- totalRevenue: sales in the last fiscal year (USD).
- costOfRevenue (USD).
- dividendsPaid: total amount of dividends paid over the last fiscal year (USD).
- CommonStockSharesOutstanding: number of total common shares outstanding.
- Dif_3-Month Treasury Bill: Secondary Market Rate: 3-month interest rate variation.
- Dif_10-Year Treasury Constant Maturity Rate: 10 years interest rate variation (%).
- 3-Month Treasury Bill: Secondary Market Rate: 3 months interest rate (%).
- 10-Year Treasury Constant Maturity Rate: 10 years interest rate (%).
- dif_unemployment_rate: difference in unemployment rate over the last fiscal year (%).
- unemployment_rate: current unemployment rate (%).
- dif_employment: difference in employment in the last fiscal year (thousands of workers).
- employment: people employed in the US (thousands of workers).
- dif_money_supply: money supply change over the last fiscal year (millions USD).
- money_supply: money supply over the last fiscal year (millions USD).
- dif_inflation_rate: difference in inflation rate in the last fiscal year (%).
- inflation_rate: current inflation rate (%).

- dif_us_gdp: difference in the GDP of the US in the last fiscal year (billions USD).
- us_gdp: US GDP for the last fiscal year (billions USD).
- dif_us_population: difference in population over the last fiscal year.
- us_population: current population.

ANNEX C

List of features obtained after preprocessing (with target encoding for *Sector*) and the correlation factor with the *target.*

```
0: ('year', -0.03160960420705554),
1: ('Ticker', 0.10133205920278085),
2: ('Industry', 0.05529133950930516),
3: ('Sector', 0.03605348627353236),
4: ('price', -0.045038514846319745),
5: ('price_adjusted', -0.06300081815924773),
6: ('price_diff1', -0.02025660842193246),
7: ('price_diff1.1', -0.08139984410950028),
8: ('market_cap', -0.011556132663420007),
9: ('eps', 0.5214423104278725),
10: ('eps_diff1', -0.006634178159408973),
11: ('eps_dif2', -0.02150081520758338),
12: ('per', 0.7553862243549447),
13: ('per_diff1', 0.01898141710607373),
14: ('per_diff2', -0.0044726665731037426),
15: ('der', 0.25954968377529875),
16: ('der_diff1', 0.021050041228322363),
17: ('der_diff2', 0.005186172261104108),
18: ('wcr', 0.32962604295566966),
19: ('wcr_diff1', -0.0030738872790048227),
20: ('wcr_diff2', -0.00016334069447912605),
 21: ('qr', 0.25818766764284695),
22: ('qr_diff1', 0.001451559553695935),
23: ('qr_diff2', -0.002370013440992083),
24: ('fcfps', 0.0009989509223767403),
25: ('fcfps_diff1', -0.029127526258520598),
26: ('fcfps_diff2', 0.03865715955290877),
27: ('pcfr', 0.02960661753462047),
28: ('pcfr_diff1', -0.04489886644363576),
29: ('pcfr_diff2', 0.026181803506354965),
30: ('totalAssets', -0.05454112291019605),
31: ('intangibleAssets', -0.038368429335338065),
32: ('otherCurrentAssets', -0.04058365889504754),
33: ('totalLiab', -0.05179261415552054),
34: ('otherCurrentLiab', -0.02318000726005412),
35: ('beginPeriodCashFlow', -0.028863291234389715),
36: ('endPeriodCashFlow', -0.021031173808498626),
37: ('totalCashFromOperatingActivities', -0.03577176873168395),
38: ('changeInCash', 0.009352110129619355),
39: ('changeInWorkingCapital', 0.010177456749309468),
40: ('freeCashFlow', -0.0065654016875797356),
41: ('incomeBeforeTax', -0.05709799262012114),
42: ('netIncome_x', -0.055291495028830615),
43: ('sellingGeneralAdministrative', -0.04590162877936608),
44: ('grossProfit', -0.05213662604647313),
45: ('ebit', -0.07025490310772327),
46: ('ebitda', -0.07452365479462746),
47: ('depreciationAndAmortization', -0.054950974429696736),
48: ('operatingIncome', -0.06651395338545688),
49: ('totalOperatingExpenses', -0.055131445296358286),
50: ('totalRevenue', -0.050899547315648726),
51: ('costOfRevenue', -0.04839214729034718),
52: ('dividendsPaid', -0.06226310824853205),
53: ('commonStockSharesOutstanding', -0.0070615141724956025),
54: ('Dif 3-Month Treasury Bill: Secondary Market Rate', -0.0690020688540976),
 55: ('Dif 10-Year Treasury Constant Maturity Rate', -0.09933016990399207),
56: ('3-Month Treasury Bill: Secondary Market Rate', -0.08035344232517067),
57: ('10-Year Treasury Constant Maturity Rate', -0.005673275782665063),
58: ('dif_unemployment_rate_proc', 0.09343368595620992),
59: ('unemployment_rate_proc', 0.12229133470724461),
60: ('dif_employment_proc', -0.08661358048060236),
61: ('employment_proc', -0.07253191357376323),
62: ('dif_money_supply_proc', -0.04414207169856849),
63: ('money_supply_proc', -0.03622668539703274),
64: ('dif_inflation_rate_proc', -0.1400865545342513),
65: ('inflation_rate_proc', -0.1440777579894728),
66: ('dif_us_gdp_proc', 0.0072646359725804),
67: ('us_gdp_proc', -0.05122013992995458),
68: ('dif_us_population_proc', 0.0072646359725804),
69: ('us_population_proc', -0.033119845546238626),
70: ('target', 1.0)
```

ANNEX D

Complete results for grid search with variability study:

N° select.	trees	min_samp.	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Avg. ret.	σ	Metric
10	50	0.001	16.98	23.71	19.25	20.89	25.09	21.18	3.28	17.90
10	50	0.005	19.17	22.09	18.45	19.10	18.54	19.47	1.50	17.97
10	50	0.010	22.43	21.82	23.42	18.50	19.48	21.13	2.06	19.07
10	50	0.050	18.85	22.48	19.31	15.81	16.47	18.59	2.64	15.94
10	100	0.001	18.73	23.39	18.60	19.24	20.75	20.14	2.01	18.13
10	100	0.005	20.47	20.65	20.24	18.79	19.28	19.89	0.81	19.08
10	100	0.010	21.33	23.73	22.05	16.92	20.72	20.95	2.52	18.43
10	100	0.050	24.79	20.61	25.05	22.06	19.72	22.44	2.41	20.03
10	200	0.001	18.11	21.39	20.91	19.33	20.92	20.13	1.37	18.76
10	200	0.005	19.81	22.93	23.00	18.19	21.89	21.16	2.11	19.06
10	200	0.010	19.39	22.44	19.76	22.96	19.56	20.82	1.73	19.09
10	200	0.050	22.41	17.40	22.56	24.34	20.30	21.40	2.66	18.75
10	500	0.001	22.60	20.60	21.17	19.79	18.32	20.49	1.59	18.90
10	500	0.005	18.77	19.23	20.13	18.67	24.68	20.29	2.52	17.78
10	500	0.010	21.36	20.68	21.66	23.62	22.20	21.90	1.10	20.80
10	500	0.050	22.88	16.52	20.77	21.86	21.64	20.73	2.47	18.26
10	1000	0.001	21.13	20.63	20.66	18.42	21.52	20.47	1.20	19.27
10	1000	0.005	18.79	18.89	20.44	19.49	24.30	20.38	2.29	18.09
10	1000	0.010	21.56	20.47	20.90	21.23	21.49	21.13	0.45	20.68
10	1000	0.050	21.80	17.84	21.63	19.72	20.90	20.38	1.64	18.74
10	2000	0.001	20.68	20.42	20.38	18.91	19.87	20.05	0.70	19.35
10	2000	0.005	17.91	21.28	22.83	20.03	25.35	21.48	2.82	18.66
10	2000	0.010	19.46	20.69	21.66	22.35	20.14	20.86	1.16	19.70
10	2000	0.050	21.35	17.80	22.00	21.20	16.24	19.72	2.54	17.18
15	50	0.001	17.30	21.64	19.65	20.29	23.14	20.40	2.19	18.21
15	50	0.005	19.61	21.59	17.12	19.55	19.52	19.48	1.58	17.90
15	50	0.010	19.73	21.93	20.26	15.99	20.47	19.67	2.22	17.46
15	50	0.050	19.04	23.58	18.30	16.02	18.58	19.10	2.76	16.35
15	100	0.001	19.83	21.29	19.17	20.44	20.63	20.27	0.81	19.47
15	100	0.005	19.85	21.39	21.79	20.99	21.19	21.04	0.73	20.31
15	100	0.010	21.18	20.97	24.19	17.98	21.67	21.20	2.21	18.99
15	100	0.050	21.73	21.89	23.72	19.91	19.51	21.35	1.70	19.65
15	200	0.001	21.26	20.03	20.73	20.80	18.57	20.28	1.05	19.23
15	200	0.005	20.50	22.37	24.30	21.82	21.16	22.03	1.45	20.58
15	200	0.010	22.41	22.51	19.44	21.03	18.38	20.75	1.82	18.93
15	200	0.050	22.41	20.59	22.01	23.74	20.02	21.76	1.48	20.27
15	500	0.001	21.79	19.25	21.55	20.25	20.64	20.70	1.03	19.67
15	500	0.005	18.20	20.00	23.57	21.17	23.14	21.22	2.23	18.99
15	500	0.010	21.60	21.49	20.93	20.47	20.08	20.91	0.65	20.26

15	500	0.050	21.74	19.34	20.07	21.32	20.21	20.54	0.98	19.56
15	1000	0.001	24.28	20.63	20.38	20.82	22.62	21.75	1.67	20.08
15	1000	0.005	19.01	22.13	21.97	21.42	22.06	21.32	1.32	20.00
15	1000	0.010	21.53	21.32	20.03	20.82	21.16	20.97	0.59	20.39
15	1000	0.050	20.84	18.67	20.07	19.92	20.20	19.94	0.79	19.15
15	2000	0.001	22.61	20.04	20.91	20.11	22.88	21.31	1.36	19.95
15	2000	0.005	21.31	21.90	21.19	20.60	22.09	21.42	0.60	20.82
15	2000	0.010	20.31	20.46	20.51	21.00	20.21	20.50	0.30	20.19
15	2000	0.050	20.64	20.80	20.58	20.22	18.84	20.22	0.80	19.42
20	50	0.001	17.30	21.76	20.53	20.12	21.27	20.20	1.74	18.46
20	50	0.005	19.09	19.99	18.11	21.40	19.20	19.56	1.23	18.33
20	50	0.010	20.70	21.25	20.33	16.27	21.06	19.92	2.07	17.85
20	50	0.050	20.78	21.92	17.56	16.00	18.43	18.94	2.40	16.54
20	100	0.001	18.17	23.32	21.77	20.21	20.16	20.73	1.93	18.80
20	100	0.005	20.19	20.25	21.69	20.63	20.99	20.75	0.62	20.14
20	100	0.010	20.92	20.38	22.89	18.25	20.83	20.65	1.66	19.00
20	100	0.050	20.52	22.21	22.21	18.56	20.24	20.75	1.53	19.21
20	200	0.001	21.29	19.03	21.02	21.29	19.08	20.34	1.18	19.16
20	200	0.005	21.12	23.31	23.79	21.74	21.59	22.31	1.16	21.14
20	200	0.010	22.04	21.82	23.08	21.03	19.07	21.41	1.50	19.91
20	200	0.050	21.63	20.19	19.67	20.88	18.97	20.27	1.03	19.24
20	500	0.001	21.70	20.49	20.83	20.76	20.56	20.87	0.48	20.38
20	500	0.005	19.68	20.81	22.65	21.14	22.02	21.26	1.14	20.12
20	500	0.010	21.18	21.24	20.93	20.51	20.90	20.95	0.29	20.66
20	500	0.050	20.29	18.93	20.95	20.24	19.77	20.04	0.75	19.29
20	1000	0.001	22.97	19.76	22.18	20.08	20.97	21.19	1.37	19.82
20	1000	0.005	20.11	22.35	21.85	21.59	22.19	21.62	0.89	20.72
20	1000	0.010	20.95	20.78	20.70	20.85	20.97	20.85	0.11	20.74
20	1000	0.050	20.27	20.16	20.49	20.22	20.23	20.27	0.13	20.15
20	2000	0.001	21.72	21.72	21.75	20.15	22.91	21.65	0.98	20.67
20	2000	0.005	21.20	22.48	21.84	20.88	22.03	21.69	0.64	21.04
20	2000	0.010	20.60	20.63	20.49	20.66	20.23	20.52	0.18	20.34
20	2000	0.050	19.45	20.34	20.39	20.32	20.13	20.13	0.39	19.74
25	50	0.001	16.51	20.66	19.86	18.71	22.15	19.58	2.12	17.45
25	50	0.005	17.90	19.60	18.95	20.37	19.39	19.24	0.91	18.33
25	50	0.010	19.83	20.39	21.05	16.81	21.32	19.88	1.81	18.07
25	50	0.050	20.12	20.71	18.08	16.93	18.05	18.78	1.58	17.20
25	100	0.001	18.60	21.74	21.85	21.19	20.76	20.83	1.32	19.51
25	100	0.005	18.75	20.44	21.66	20.04	22.64	20.71	1.50	19.21
25	100	0.010	21.25	20.41	23.11	19.91	20.74	21.09	1.23	19.85
25	100	0.050	19.58	21.76	21.82	20.58	19.72	20.69	1.08	19.62
25	200	0.001	22.16	18.68	22.15	20.77	20.86	20.92	1.42	19.50
25	200	0.005	20.93	22.52	22.95	20.26	20.84	21.50	1.17	20.33
25	200	0.010	22.05	22.07	22.68	20.45	22.09	21.87	0.84	21.03

25	200	0.050	21.19	19.79	19.66	20.10	18.29	19.80	1.04	18.77
25	500	0.001	21.38	19.44	21.26	20.41	21.37	20.77	0.85	19.93
25	500	0.005	19.48	21.59	22.38	20.47	21.05	21.00	1.10	19.89
25	500	0.010	21.03	21.67	21.41	20.95	21.03	21.22	0.31	20.91
25	500	0.050	19.62	20.09	21.12	19.14	20.46	20.08	0.76	19.32
25	1000	0.001	22.03	21.44	23.25	20.94	21.55	21.84	0.88	20.96
25	1000	0.005	21.30	20.95	21.80	20.99	21.65	21.34	0.38	20.96
25	1000	0.010	21.10	19.50	20.66	19.77	21.34	20.47	0.81	19.66
25	1000	0.050	20.12	20.35	19.96	20.18	20.12	20.15	0.14	20.01
25	2000	0.001	22.89	22.58	22.35	22.61	21.64	22.41	0.47	21.94
25	2000	0.005	21.63	20.94	21.92	21.94	21.85	21.66	0.42	21.24
25	2000	0.010	20.19	19.56	21.10	19.96	20.86	20.33	0.64	19.70
25	2000	0.050	18.63	19.82	18.94	18.96	19.70	19.21	0.52	18.69

ANNEX E

Companies selected by the final model though the years are the following:

Year	Ticker	Return (%)
2000	EQR	12.94
2000	FRT	29.35
2000	USB	-6.58
2000	CINF	-0.83
2000	AFL	-27.80
2000	AME	27.25
2000	CL	-2.60
2000	PNC	-15.80
2000	ADM	19.02
2000	MET	-11.69
2000	FITB	7.53
2000	PCAR	44.97
2000	UDR	52.50
2000	EFX	34.39
2000	GL	0.94
2000	PSA	54.13
2000	AXP	-36.77
2000	ALL	-21.04
2000	SCHW	-46.99
2000	BMY	-19.43
2000	KMB	-11.80
2000	ARE	17.14
2000	LNC	2.57
2000	EXPD	9.13
2000	CPT	22.11
2001	IDXX	11.63
2001	CNC	66.53
2001	CRL	19.65
2001	MAA	2.96
2001	WEC	13.44
2001	HSY	0.87
2001	STLD	16.07
2001	HAL	36.98
2001	PCAR	11.16
2001	TXT	10.17
2001	EXPD	20.72
2001	O	30.35
2001	KIM	1.44
2001	GL	-3.45
2001	MLM	-30.61
2001	VFC	-3.95
2001	CSGP	-13.16
2001	ETN	6.55
2001	XRAY	12.87
2001	ACGL	29.57
2001	ARE	5.95
2001	TRMB	-18.29
2001	EQR	-7.24
2001	CINF	3.07
2002	UDR	28.51
2002	PEAK	38.12
2002	BRO	-1.37
2002	AME	27.38
2002	IRM	23.32
2002	IEX	29.18
2002	WAB	23.34
2002	CTRA	17.45
2002	KIM	50.17
2002	WELL	41.90
2002	CME	63.86
2002	FAST	29.53
2002	ECL	13.40
2002	WEC	38.47
2002	EXPD	14.32
2002	CHD	34.66
2002	ACGL	24.18
2002	PCAR	81.44
2002	BWA	66.97
2002	PKG	20.35
2002	ODFL	90.77

2002	WYNN	102.21
2002	MCO	39.38
2002	EW	15.57
2002	BALL	17.28
2003	CHRW	44.40
2003	CHD	22.40
2003	PEAK	19.89
2003	ORLY	8.72
2003	AOS	-12.98
2003	XRAY	23.52
2003	UHS	-12.30
2003	VFC	32.64
2003	PKG	11.26
2003	MNST	295.02
2003	EOG	58.42
2003	CNC	93.74
2003	ECL	29.10
2003	JBHT	58.45
2003	STLD	73.02
2003	LKQ	10.96
2003	WELL	12.07
2003	IEX	48.50
2003	TROW	40.86
2003	ACGL	-0.29
2003	PLD	30.18
2003	EBAY	94.99
2003	MAR	34.69
2003	CSGP	8.22
2003	HAS	-8.17
2004	CHD	3.06
2004	AME	25.08
2004	CNC	-9.30
2004	WAB	31.43
2004	AJG	1.96
2004	IRM	42.31
2004	HSIC	28.10
2004	WST	6.66
2004	REG	13.82
2004	SHW	2.10
2004	PPL	17.88
2004	PNW	0.42
2004	CHRW	40.47
2004	RSG	14.28
2004	AEE	10.56
2004	BRO	43.32
2004	ORLY	44.14
2004	PEAK	1.17
2004	CBRE	83.24
2004	EXPD	27.59
2004	CINF	9.77
2004	DTE	5.08
2004	GL	-1.61
2004	DPZ	40.69
2004	PEG	43.83
2005	CHD	29.17
2005	AME	11.66
2005	FRT	39.58
2005	ICE	200.26
2005	PEAK	46.59
2005	ODFL	-0.86
2005	CTRA	38.20
2005	EXPD	23.72
2005	VTR	34.41
2005	UDR	44.74
2005	ROL	10.94
2005	TSCO	-15.02
2005	EXR	28.03
2005	AJG	1.14
2005	EQIX	89.53
2005	NSC	16.79
2005	WST	105.71
2005	PLD	27.82
2005	XEL	29.70
2005	PNW	24.15
2005	ATVI	28.17
2005	ECL	29.37
2005	MAR	39.16
2005	CNC	-3.09

2005	SEE	17.97
2006	PNR	11.23
2006	LKQ	79.38
2006	BRO	-15.73
2006	CHD	31.69
2006	PKG	31.38
2006	HSIC	21.15
2006	AME	47.68
2006	VTR	16.48
2006	IEX	16.67
2006	ROL	35.97
2006	NWL	-7.75
2006	YUM	30.70
2006	ODFL	-7.46
2006	PEAK	-4.11
2006	CHRW	28.54
2006	CSGP	-10.96
2006	AOS	1.19
2006	RHI	-29.12
2006	CMI	105.01
2006	AEE	4.92
2006	DLR	12.41
2006	FRT	1.59
2006	GWW	26.49
2006	ITW	18.94
2006	PPL	49.80
2007	ROL	-15.00
2007	CHD	-2.98
2007	TSCO	-2.18
2007	PWR	-31.03
2007	FAST	-17.51
2007	AAP	-16.56
2007	ODFL	0.58
2007	CSGP	-34.49
2007	K	-19.01
2007	EXPD	-29.51
2007	CTRA	-32.56
2007	DOV	-33.68
2007	ALB	-53.67
2007	DHR	-39.29
2007	RHI	-25.30
2007	WY	-46.84
2007	NEE	-29.56
2007	POOL	-17.57
2007	WBD	-53.10
2007	AOS	-16.86
2007	GPC	-18.90
2007	PPG	-37.34
2007	CL	-17.95
2007	MAA	-17.10
2007	NDAQ	-49.69
2008	ODFL	28.72
2008	EW	68.04
2008	ROL	16.85
2008	WAB	2.03
2008	EXR	31.69
2008	DLR	71.98
2008	AJG	-2.48
2008	TSCO	37.44
2008	LKQ	83.10
2008	IFF	45.38
2008	DPZ	92.38
2008	CHD	13.22
2008	JBHT	34.94
2008	NDAQ	-15.39
2008	MAR	60.14
2008	PXD	167.30
2008	ACGL	5.45
2008	FAST	20.25
2008	MPWR	125.97
2008	ITW	52.24
2008	IEX	42.46
2008	MKTX	90.70
2008	POOL	14.00
2008	WST	9.92
2008	CDNS	77.34
2009	PNR	14.71
2009	WST	5.06

2009	CSGP	34.90
2009	RHI	19.29
2009	ODFL	51.41
2009	CNC	22.87
2009	LKQ	19.30
2009	PWR	-2.82
2009	CHD	14.21
2009	DLR	9.68
2009	EW	79.93
2009	REG	27.94
2009	CHRW	36.46
2009	ROL	54.90
2009	SHW	32.73
2009	FMC	43.41
2009	CPT	34.89
2009	BRO	34.67
2009	EFX	17.57
2009	VRSK	12.87
2009	TDY	16.40
2009	LNT	28.17
2009	ZBRA	38.79
2009	XRAY	-2.64
2009	AWK	17.75
2010	WAB	30.63
2010	IEX	-6.22
2010	WST	-6.61
2010	FTNT	40.06
2010	NWL	-11.63
2010	CMS	17.89
2010	LNT	20.85
2010	ODFL	27.92
2010	EVRG	15.52
2010	CINF	-0.47
2010	AEE	20.07
2010	AJG	16.86
2010	PLD	-6.42
2010	BRO	-6.86
2010	FRT	20.89
2010	SEE	-27.90
2010	FLT	-1.51
2010	PNR	-0.64
2010	AWK	29.10
2010	JBHT	13.32
2010	MCO	29.41
2010	IDXX	7.95
2010	FAST	46.24
2010	EFX	8.81
2010	MOH	23.71
2011	WEC	14.58
2011	O	23.77
2011	PNW	14.15
2011	NI	12.33
2011	TYL	55.88
2011	FAST	6.92
2011	CMS	20.09
2011	LNT	8.47
2011	NDAQ	0.37
2011	DVA	45.11
2011	CSGP	31.85
2011	BRO	20.41
2011	ED	-3.06
2011	PLD	30.74
2011	SBAC	68.50
2011	COP	11.93
2011	XEL	5.80
2011	PODD	17.30
2011	CCI	60.48
2011	ANSS	13.85
2011	MKTX	20.33
2011	PEP	10.40
2011	BMY	-0.46
2011	CVS	24.76
2011	CVX	7.29
2012	FMC	32.26
2012	PODD	70.93
2012	WST	80.70
2012	O	-2.05
2012	MAA	-0.20

2012	PEAK	-15.37
2012	ANSS	27.39
2012	MPWR	57.09
2012	GNRC	86.21
2012	TSCO	72.11
2012	ROL	33.81
2012	SWK	13.88
2012	AMT	4.17
2012	CCI	5.25
2012	DXCM	152.90
2012	JBHT	31.53
2012	AME	35.37
2012	IVZ	42.57
2012	SYK	35.59
2012	HSIC	40.35
2012	WY	13.11
2012	VRSK	31.48
2012	TRMB	13.27
2012	BRO	19.42
2012	EFX	27.39
2013	FAST	1.23
2013	IVZ	16.11
2013	PWR	-6.17
2013	AXON	49.64
2013	RSG	20.98
2013	BWA	2.27
2013	ROL	14.21
2013	ALLE	27.25
2013	CMS	33.91
2013	ANSS	-3.72
2013	EFX	20.92
2013	NOW	23.53
2013	AVY	5.79
2013	AOS	4.16
2013	HSIC	19.93
2013	VTRS	32.19
2013	EPAM	42.30
2013	TDY	13.74
2013	MPWR	49.28
2013	CRL	20.82
2013	NI	35.10
2013	AWK	30.62
2013	CHD	19.63
2013	AON	17.01
2013	FIS	22.36
2014	MPWR	34.48
2014	CMS	7.36
2014	EVRG	8.48
2014	XRAY	13.20
2014	CNP	-21.84
2014	AJG	-8.25
2014	MMC	-1.17
2014	ACGL	21.39
2014	CDW	26.14
2014	NWL	24.68
2014	NI	20.61
2014	WM	9.35
2014	CHD	11.86
2014	CINF	20.37
2014	K	11.12
2014	ZBH	-10.04
2014	IPG	16.86
2014	MAS	35.28
2014	DPZ	16.98
2014	REG	9.01
2014	RSG	13.99
2014	JNPR	32.39
2014	CBRE	4.60
2014	ANET	10.48
2015	WST	36.21
2015	UDR	-2.16
2015	DPZ	50.43
2015	ANET	30.37
2015	EXR	-12.49
2015	AOS	27.69
2015	CBRE	-10.40
2015	ETSY	35.38
2015	EPAM	-21.36

2015	FRT	-1.61
2015	IPG	5.56
2015	CINF	31.75
2015	MPWR	27.68
2015	ANSS	2.05
2015	IT	13.44
2015	AVB	-2.26
2015	MSCI	13.58
2015	DXCM	-24.00
2015	CRL	-4.34
2015	ROL	28.03
2015	IQV	12.06
2015	INCY	-7.68
2015	ACGL	21.23
2015	RE	19.21
2015	MLM	56.36
2016	PGR	61.79
2016	BRO	16.99
2016	CMS	21.69
2016	RSG	18.51
2016	PNW	18.98
2016	ES	22.10
2016	AWK	27.17
2016	AJG	29.83
2016	CHD	11.81
2016	AEE	22.36
2016	PEG	25.37
2016	BK	15.20
2016	GL	23.25
2016	ODFL	48.13
2016	NTRS	12.91
2016	EMN	24.09
2016	DOV	32.82
2016	MOS	-14.30
2016	SWK	44.97
2016	AON	23.20
2016	PPL	2.51
2016	SCHW	31.02
2016	LKQ	24.51
2016	NI	22.45
2016	FITB	15.69
2017	RHI	6.34
2017	POOL	16.52
2017	FAST	2.50
2017	FTV	-5.39
2017	IRM	-7.62
2017	IR	-30.69
2017	EFX	-17.48
2017	LUV	-22.20
2017	ETSY	156.86
2017	HBAN	-12.59
2017	ROL	21.43
2017	VTR	5.08
2017	TER	-21.81
2017	AOS	-28.81
2017	JNPR	-1.29
2017	BK	-11.26
2017	O	20.61
2017	CMS	9.27
2017	AWK	5.57
2017	AMP	-33.16
2017	CHD	37.57
2017	IEX	-0.63
2017	DXCM	106.53
2017	LIN	4.50
2018	CDAY	79.88
2018	TYL	58.64
2018	AXON	66.10
2018	MKTX	75.48
2018	ETSY	-16.32
2018	DXCM	81.99
2018	MPWR	47.46
2018	PAYC	109.69
2018	EXR	14.43
2018	NOW	55.74
2018	AWK	32.17
2018	MRNA	18.09
2018	EVRG	11.44

2018	EPAM	76.69
2018	VRSN	25.89
2018	TSCO	9.90
2018	IDXX	34.34
2018	EQR	22.77
2018	RHI	8.27
2018	ROL	-8.26
2018	PODD	122.73
2018	ENPH	370.01
2018	ES	26.00
2018	WEC	29.24
2018	FAST	39.78
2019	EXR	11.98
2019	MPWR	96.31
2019	O	-13.70
2019	WELL	-17.18
2019	NDAQ	23.11
2019	CDAY	57.02
2019	UDR	-14.48
2019	MAA	-2.49
2019	IT	0.41
2019	CPT	-4.71
2019	CRL	64.46
2019	TYL	50.57
2019	SBAC	18.59
2019	REG	-22.01
2019	XYL	29.83
2019	BRO	18.68
2019	PWR	73.21
2019	DPZ	35.21
2019	FAST	39.02
2019	MRNA	620.28
2019	IRM	-0.38
2019	VRSK	35.40
2019	AWK	25.51
2019	MKTX	49.87
2019	AJG	30.77
2020	CHRW	11.84
2020	ROL	-14.34
2020	HSIC	10.01
2020	AOS	51.76
2020	TRMB	33.49
2020	EMN	16.94
2020	VTR	3.00
2020	IRM	80.34
2020	JNPR	55.80
2020	CBRE	61.66
2020	KIM	62.49
2020	SPG	85.88
2020	REG	61.72
2020	CPT	80.50
2020	GRMN	14.81
2020	PNR	41.35
2020	VFC	-13.09
2020	WRB	24.84
2020	PWR	60.58
2020	YUM	26.03
2020	RSG	44.71
2020	DOV	42.24
2020	BRO	48.46
2020	UDR	57.00
2020	TXT	57.17
2021	BRO	-14.19
2021	XYL	-5.60
2021	EXR	-26.25
2021	EW	-38.70
2021	VICI	23.57
2021	TRMB	-37.63
2021	GRMN	-28.83
2021	DXCM	-15.87
2021	VRSN	-17.77
2021	CCI	-27.58
2021	ROL	17.93
2021	AKAM	-23.80
2021	CHD	-15.68
2021	PLD	-26.74
2021	FAST	-19.60
2021	WM	1.26

2021	TER	-43.05
2021	WEC	4.45
2021	AME	-0.69
2021	VRSK	-20.06
2021	VMC	-10.75
2021	MAS	-27.46
2021	TYL	-37.37
2021	SYK	-4.49
2021	WAT	-2.29