NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

PORTFOLIO OPTIMIZATION: FROM MARKOWITZ TO MACHINE LEARNING

Mariana Serrano Lopes da Bernarda

Project Work presented as partial requirement for obtaining the Master's degree in Statistics and Information Management, Specialization in Risk Analysis and Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PORTFOLIO OPTIMIZATION: FROM MARKOWITZ TO MACHINE LEARNING

by

Mariana Serrano Lopes da Bernarda

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Risk Analysis and Management

Advisor / Co Advisor: Rui Gonçalves

November 2021

ABSTRACT

In the past few decades, substantial progress has been made in portfolio optimization, especially with the emergence of machine learning. Therefore, it is essential to find the models that not only achieve the best results but also simplify the process. This project aims to demonstrate that to achieve optimal portfolios cannot be based only on traditional statistical methods. Therefore the Random Forest regression model, a machine learning model, was chosen to predict stock prices to complement the Markowitz model, a classical portfolio selection model.

To evaluate the efficacy of the modified model compared to the classical model the following methodology was adopted: data was collected (from 2012 to 2019 from 10 companies and it was divided in 15 periods) and treated; some common technical indicators were extracted; one stock price was predicted per period; expected returns and partially estimated volatility were derived from the predictions and introduced in the classical model; 15 portfolios were constructed by each model; and finally, a performance analysis was conducted. The results obtained show that the 1-day predictions were quite accurate, almost 90%, and the modified model's portfolios' outperformed the classical model's portfolios for most periods analyzed.

KEYWORDS

Portfolio Optimization; Machine Learning; Random Forest; Markowitz; Sharpe Ratio

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
AXP	American Express Co
CAL	Capital Allocation Line
CTXS	Citrix Systems
DIS	The Walt Disney Company
ECL	Ecolab Inc.
EMH	Efficient Market Hypothesis
F	Ford Motor Company
FDX	FedEx Corporation
GDA	Gaussian Discriminant Analysis
MACD	Moving Average Convergence Divergence
MAE	Mean Absolute Error
ML	Machine Leaning
MO	Altria Group Inc
MPT	Modern Portfolio Theory
MSE	Mean Squared Error
OBV	On Balance Volume
OXY	Occidental Petroleum
PROC	Price Rate of Change
QDA	Quadratic Discriminant Analysis
RF	Random Forest model
RSI	Relative Strength Index
S&P 500	Standard & Poor's 500 Index
SVM	Support Vector Machines
UNH	United Health Group Inc.
XEL	Xcel Energy Inc

1. INTRODUCTION

When we face countless choices, we tend not to make a decision. Portfolio managers constantly go through complicated decision-making processes such as the selection of stocks to invest in, the weighing of each stock, and among others. Traditional statistics do not facilitate these processes because of the non-stationary and non-linearity characteristics of the stock market (Zhang et al., 2016). Therefore researchers started using machine learning (ML) to predict the stock market and improve results.

ML is defined by Alpaydin (2010) as "programming computers to optimize a performance criterion using example data or past experience". However, it is not just algorithms; it is also part of artificial intelligence. In a changing environment, a system has to be able to learn and adapt to provide solutions.

ML can be applied in various fields, namely in portfolio optimization which is defined as the process of choosing the best assets out of the considered and adjusting each of their weights, according to an objective. Portfolio optimization has made substantial progress since Modern Portfolio Theory (MPT), developed by Markowitz (1952), and according to the theory, an efficient frontier of optimal portfolios can be constructed offering the highest level of return for a given level of risk.

The portfolio selection process consists of two steps: the first step is to analyze the historical data to have a sense of the assets future behavior and the second step is to construct the portfolio based on the first step's insights (Markowitz et al., 1952). This project attempts to use ML to predict stock prices in the first step of the portfolio selection process.

There are various ML techniques, namely supervised, unsupervised, and reinforcement learning. Supervised learning is a class of algorithms that learn from a training dataset, models such as linear regression, random forest (RF), and neural networks. Unsupervised learning finds and analyzes hidden patterns in data; some common algorithms are clustering and anomaly detection. Reinforcement learning is a group of algorithms which are trained using a system of reward and punishment, one of the most known models is Q-learning (Marsland, 2015).

This project aims to demonstrate that to achieve optimal portfolios cannot be based only on traditional statistical methods. Therefore RF regression model was chosen to predict stock prices and derive from them expected returns and partially estimated volatility to complement the classical Markowitz model. The choice of using this ML model was based on the fact that it is one of the most common ML ensemble models. Many of the published studies on the RF algorithm attest to it having very high average performance, this will be analyzed further to verify if the project's predictions are in line with the literature. This modified model and the classical model will derive 15 portfolios each and will be compared using performance measures later in the project. The study objectives are summarized in the table below as well as the respective references:

Objectives	Research Questions	References
Compare portfolios obtained from the classical Markowitz model and the extended version where RF regression algorithm is used to predict stock prices and derive from them expected returns and partially estimated volatility	Other than RF regression algorithms, what are other existing models?	A comprehensive literature review will be provided
	Will the modified version of the classical model be more efficient by producing better results than the classical model?	Classical Markowitz model (Markowitz, 1952)
		RF regression algorithm (Pedregosa et al., 2011) with technical indicators (Khaidem et al., 2016)
		Accuracy measures (Hyndman and Athanasopoulos, 2018) will be used to verify the accuracy of the predictions and performance measures will be used to compare the portfolios selected by both models
	Is the machine learning model used sufficient or are there any needed improvements that can be suggested?	Some suggestions to implement in the model could include other technical indicators as applied in Xinjie (2014)
		Other suggestions could include external data that influence the stock price as tested in Li et al. (2014)
	Based on results, how is machine learning able to help portfolio optimization?	Bartram et al. (2020)

Table 1 - Proposed Objectives

The processes explored in this paper tried to predict the market using it as assistance to human decision-making processes. The results and conclusions could not only be beneficial to industry practitioners but also bring economic and social benefits if people recognized that machine learning models could be an asset if they were better studied and implemented.

This project is organized as follows: (i) in section 2 a comprehensive literature review is provided, (ii) in section 3 the processes for data preprocessing are described in detail namely data collection, data treatment and feature extraction as well as the model implementation (where the chosen models are explained in detail and what was done to obtain the portfolios), (iii) in section 4 the results are provided as well as a performance analysis, and (iv) in section 5 the discussion and the conclusions are provided.

2. LITERATURE REVIEW

The portfolio optimization problem consists of many unknown variables having to be predicted, for example expected returns, and this implies a high sensitivity to the precision of the prediction methodologies (Tadlaoui, 2018). Since Markowitz introduced the MPT in 1952, this problem has been widely studied by many researchers.

The main ideas of Markowitz's theory are that return and risk should be analyzed together and that a portfolio should be diversified. When constructing a portfolio with correlated assets the losses of some assets can be offset by the gains of other assets. Markowitz's theory has been extended by many researchers to improve its main limitations, such as the high sensitivity to historical prices and not being able to contain investors' views (Zhang et al., 2018).

The Efficient Market Hypothesis (EMH), developed by Fama (1970), suggests that it is impossible to outperform the market by trying to predict future stock prices because prices fully reflect all relevant information. Many researchers that disagreed with this theory believed that stock prices were partially predictable and started using algorithms that are able to model the stock market (Malkiel, 2003).

In the past decades, more ML predicting algorithms started to be applied in finance, such as Artificial Neural Network (ANN), Support Vector Machines (SVM), RF, and many others (Prado, 2018). Each algorithm has suffered and still is suffering many modifications over the years; therefore there are innumerous studies on many variations of models.

Some researchers analyzed algorithms applied to the next-day model, which forecasts the outcome of the share price on the next day, and to the long-term model, which forecasts the outcome of the share price for the next n days. Dai and Zhang (2013) analyzed Logistic Regression, Gaussian Discriminant Analysis (GDA), Quadratic Discriminant Analysis (QDA), and SVM algorithms. The data used contained daily stock prices from the company 3M Stock from 01/09/2008 to 11/08/2013 and 16 features were extracted. The results showed that the long-term model presented better results with the SVM algorithm attaining a success rate of 79.3%.

Researchers also applied various features to the algorithms trying to improve their accuracy. Xinjie (2014) used three stocks with time span available from 04/01/2010 to 10/12/2014 and used an extremely randomized tree algorithm (Geurts and Louppe, 2011) to select from 84 features the top 30% of features and introduce them to the SVM algorithm. Technical indicators included were the Relative Strength Index (RSI), the Rate of Change, and among others. The results showed above 70% accurate prediction.

Other researchers included in the models data that influenced the estimated variables. Li et al. (2014) analyzed linear and SVM models and took into consideration how stock prices can be influenced by external conditions. The external conditions considered were daily quotes of commodity future contracts, 2 foreign currencies (EUR, JPY) and 1 interest rate. In addition to that, daily US stocks data was also collected; the data was from 01/01/2000 to 10/11/2014. The features constructed were direct (from the stocks data) and indirect (from external factors); these features were normalized and centralized. The results indicated that out of the models analyzed, the best was the logistic regression with a success rate of 56.65%.

Decision tree models have very high variance and low bias. Despite the RF model being an ensemble of various decision trees it does not have the problem of high variance because it trains the decision trees on distinct subspaces risking slightly increased bias. The RF model, as stated by Biau and Scornet (2016), was originally developed by Breiman (2001) but has experienced many extensions since then.

Some extensions to the RF model include changing tree weights since the original model's final prediction is the average of the aggregated predictions of the individual trees. Winham et al. (2013) increased the weights of better performing trees to increase accuracy and this model outperformed the traditional RF model. Bernard et al. (2012) proposed a similar model, one that would grow only trees that complement the existing trees in the ensemble to avoid the forest performance to decrease and it also outperformed the traditional RF model.

The original RF model is an offline algorithm, for there to be an output it is necessary to input a whole dataset. Unlike online algorithms, they do not require inputting a whole dataset at once. These models are useful when data is produced over time and has to be inputted into the model quickly. Lakshminarayanan et al. (2014) proposed a model where the trees grow online and achieved competitive predictive performance.

Ishwaran et al. (2008) introduced an extension of the classical RF model, random survival forests model. Survival analysis attempts to analyse duration of time until one or more events happen and very often there is incomplete data. The created model includes new splitting rules for growing trees, a new missing data algorithm for imputing missing data and a conservation-of-events principle. This model was consistently more accurate than competing models.

Khaidem et al. (2016) were among the few researchers that exponentially smoothed the data before extracting features to input into the chosen models. They analyzed the RF model and compared it to various other models. The results of the RF were well above 80% surpassing many of the models mentioned in this section. Basak et al. (2019) also exponentially smoothed the data and applied various technical indicators to tree based classifiers namely, RF and XGBoost. The XGBoost outperformed the RF for longer prediction window but RF had higher success rate with shorter prediction window.

In the stock market prediction problem, the ensemble learning models are very common models that have very high average performances. Therefore, this project uses an ensemble learning model, RF regression algorithm (Pedregosa et al., 2011) with technical indicators (Khaidem et al., 2016), to predict stock prices and try to demonstrate that results are improved when classical models are paired up with machine learning models.

3. METHODOLOGY

The project followed the methodology shown in figure 3. For the modified version of the Markowitz model, the dataset collected was exponentially smoothed which is suitable for data forecasting with no clear trend or seasonal pattern. Then features/ technical indicators were extracted from the dataset and implemented in the model.

For the classical Markowitz model, the dataset was not treated nor were technical indicators extracted. Both models were then implemented and portfolios were constructed. The final step was an analysis of the accuracy of the predictions and an analysis of the performance of the portfolios of each model. The following sections will include a detailed description of all these steps.

Figure 3 – Project Methodology

3.1. DATA COLLECTION

The dataset includes historical prices of stocks between 2012 and 2019 from 10 companies extracted from Yahoo Finance website. These companies were chosen at random from the S&P500 index and all have an inception date before 2000. Some of these companies are from the Communication Services sector (Disney (DIS)), Energy sector (Occidental (OXY)), Industrials sector (FedEx (FDX)), among others. As we can see in the table below, the selected companies are from different sectors which are crucial to ensure diversification and effectiveness of the algorithms.

Symbol	Security	Sector
MO	Altria Group Inc	Consumer Staples
AXP	American Express Co	Financials
CTXS	Citrix Systems	Information Technology
DIS	The Walt Disney Company	Communication Services
ECL	Ecolab Inc.	Materials
FDX	FedEx Corporation	Industrials
F	Ford Motor Company	Consumer Discretionary
OXY	Occidental Petroleum	Energy
UNH	United Health Group Inc.	Health Care
XEL	Xcel Energy Inc	Utilities

Table 3.1 – Selected Companies

The raw values from the dataset considered include 7 columns namely: date, open price, highest price, lowest price, close price, adjusted close price (which is an adjustment to the close price that takes into account any corporate actions), and transaction volume (which is the total number of shares transacted during the day). For this project, the adjusted close price was used for all the calculations and predictions.

3.2. DATA TREATMENT

The selected data was not directly inputted in the RF model to make predictions; data treatment was conducted to avoid discrepancies. Exponential smoothing was applied to attribute larger weights to recent data and exponentially reduce weights of older data. This data treatment method was used to reduce the effects of jumps and abrupt changes in the dataset.

Given a time series $P = (P_t)_{t\geq 0}$, the exponential smoothed version $\hat{P} = (\hat{P}_t)_{t\geq 0}$ can be recursively calculated as (Hyndman et al., 2018)):

$$\hat{P}_0 = P_0$$

$$\hat{P}_{t+1} = \alpha P_{t+1} + (1-\alpha)\hat{P}_t$$

Where α is the smoothing constant, a value from 0 to 1. Higher values of α reduce the level of smoothing. This smoothing removes the random changes in the historical data, enabling the model to easily recognize long-term price trends in the dataset. For this project, α = 0.2 was considered since Ravinder (2013) suggested a smoothing factor below 0.5. The graph below shows the original and the smoothed prices of Altria in period 1. The same methodology was applied to the other periods as well as to the other companies.

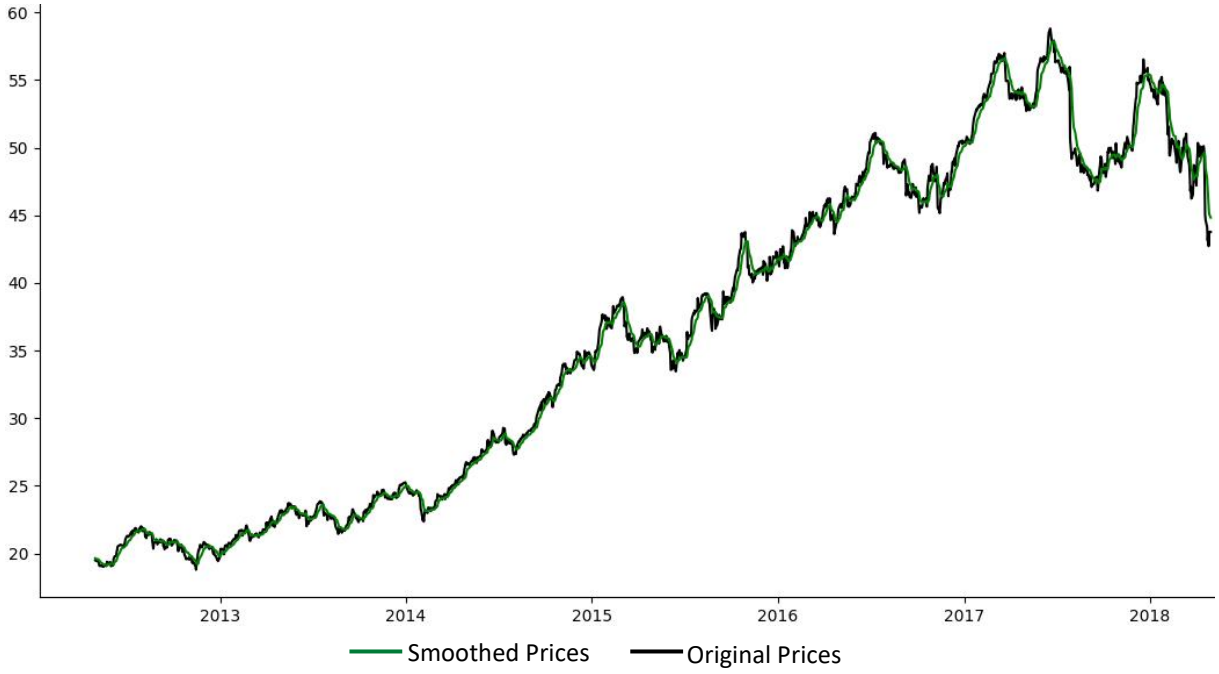

Figure 3.2 – Altria's exponential smoothed prices for period 1

3.3. FEATURE EXTRACTION

Features or technical indicators are mathematical calculations based on historic price, volume, or (in terms of a futures contract) open interest information that help investors understand price movements. The features that were extracted from the smoothed data were the same applied in Khaidem et al. (2016) and were inputted in this project's RF model. They are listed below:

- **Moving Average Convergence Divergence**

The moving average convergence divergence (MACD) (Appel and Dobson, 2008) is a momentum indicator that shows the relationship of two moving averages of prices. The MACD is calculated as follows: the 26-day exponential moving average (EMA) is subtracted from the 12-day EMA. The 9-day EMA of the MACD is the signal line, the baseline for the buy and sell signals. The formula for MACD is:

$$MACD = EMA_{12} - EMA_{26}$$

$$SignalLine = EMA_9(MACD)$$

- **On Balance Volume**

On balance volume (OBV) (Granville, 1976) is a technical indicator that predicts changes in stock prices based on the cumulative volume. When the price goes down, the volume traded is subtracted; and when the price goes up the volume trade is accumulated. The formula for OBV is:

$$OBV(t) = OBV(t-1) + \begin{cases} Volume(t) & if\ P_t > P_{t-1} \\ 0 & if\ P_t = P_{t-1} \\ -Volume(t) & if\ P_t < P_{t-1} \end{cases}$$

Where P(t) is the closing price at time t.

- **Price Rate of Change**

The price rate of change (PROC) is an indicator that measures the percentage change in price between the current price and the price over the period considered (Khaidem et al, 2016). The formula for PROC is:

$$PROC_t = \frac{P_t - P_{t-n}}{P_{t-n}}$$

Where,

P(t) is the closing price at time t

P(t-n) is the closing price n periods before time t

- **Relative Strength Index**

The relative strength index (RSI) (Wilder Jr, 1978) is an indicator that evaluates the magnitude of recent price changes to determine if a stock is overbought or oversold. RSI ranges from 0 to 100 and normally, when the RSI is below 30 the stock is oversold and when the RSI is above 70 the stock is overbought. The formula for RSI is:

$$RSI = 100 - \frac{100}{1 + RS}$$

$$RS = \frac{Average\ gain\ over\ past\ 14\ days}{Average\ loss\ over\ past\ 14\ days}$$

- **Stochastic Oscillator**

The stochastic oscillator (Lane, 1984) is an indicator that compares a particular price of a stock to a range of prices over a period of time. The formula for stochastic oscillator is:

$$\%K = \left(\frac{C - L_{14}}{H_{14} - L_{14}}\right) \times 100$$

Where,

C=current closing price

L_{14}=lowest price over the past 14 days

H_{14}=highest price over the past 14 days

- **Williams Percent Range**

The Williams percent range indicator, designed by Larry Williams, is similar to the stochastic oscillator indicator. It compares a stock's price to the high-low range over a specific period, normally 14 days. It ranges from -100 to 0 and normally, it indicates a sell signal when it is above -20 and it indicates a buy signal when it is below -80 (Basak et al., 2019). The formula for Williams percentage range is:

$$\%R = \left(\frac{H_{14} - C}{H_{14} - L_{14}}\right) \times -100$$

Where,

C=current closing price

L_{14}=lowest price over the past 14 days

H_{14}=highest price over the past 14 days

3.4. Model Implementation

3.4.1. Random Forest Model

As mentioned above, the ML model used to predict stock prices was from Scikit-Learn Package (Pedregosa et al., 2011), more precisely the sklearn.ensemble.RandomForestRegressor, following the methodology in Breiman (2001), and some technical indicators were added to the model (Khaidem et al., 2016). Portfolio optimization depends mainly on personal experience and knowledge of the trader and RF model imitates the human thought process. It is formed of various Decision Trees and it learns from a training dataset applying its training on the estimation of stock prices.

The model is comprised of three parts namely, the training set, the testing set, and the simulation. The training set is the actual dataset used to train the model, after the model is trained the testing set is used for performance evaluation, and then the simulation/ prediction is conducted. In this project, 15 predictions were performed and all adopted the following methodology: the training set includes the first 4/5 of the dataset, the testing set includes the remaining 1/5, and the forecast is of the next trading day, as seen in the diagram below.

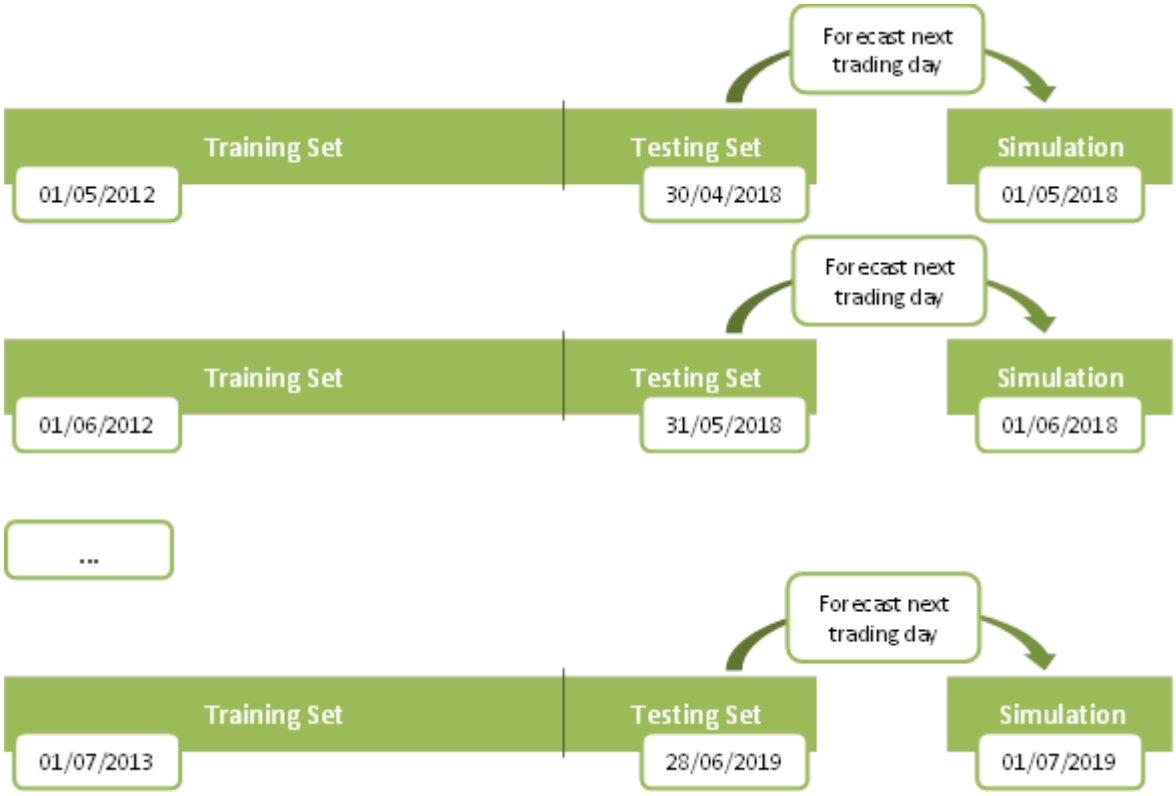

Figure 3.4.1 – Prediction Metholodogy

As mentioned before, the RF is an ensemble of Decision Trees that are trained on distinct subspaces which means that the trees cannot see the whole dataset. Data samples are randomly distributed with replacement, known as bootstrapping (Breiman et al., 1984), meaning that some data samples will be used many times in a single tree. The aim of this technique is to reduce high bias and high variance of the entire forest.

In this model, there are 100 trees in the forest; and in each tree, each node is repeatedly split into subsets. The split is done by asking a question on a characteristic. The mean squared error (MSE) is an impurity measure that is used as splitting criterion; it measures the quality of the split (Pedregosa et al., 2011):

$$MSE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|^2$$

Where,

n = is the number of items
y_j = is the true value
$\hat{y}_j$ = is the prediction

Another ensemble technique that is present in the RF algorithm is bagging (Breiman, 1996) which is an average of aggregated predictions from all the trees in the forest, also known as the final prediction of the model. This technique can drastically reduce variance leading to improved predictions.

3.4.2. Markowitz Portfolio Selection Model

As mentioned above, the classical model used was the Markowitz model that will be compared further in the thesis to a modified version that includes the prediction of stock prices using the RF model. According to Markowitz's theory, an optimal portfolio is one that achieves minimal volatility with an acceptable expected rate of return.

The Markowitz theory has the following assumptions (Markowitz, 1952): investors are rational and want to maximize their utility, investors have access to all information needed, markets are efficient, investors are risk-averse and base their decisions accordingly, and for a given level of risk, investors prefer higher returns to lower returns. Some of these assumptions are unrealistic because not all investors have the same investment strategies and not all are risk-averse.

Following the Markowitz theory, the raw dataset was used to calculate each asset's return and volatility. For the modified version, the estimated prices were used to obtain the assets' expected return and partially estimated volatility. The volatility or risk was derived from the standard deviation of the prices. The correlation between assets is also relevant to construct a portfolio because when assets are less correlated a portfolio is more diversified, leading to higher expected returns and lower risks.

The formulas for return are given by (Bodie et al., 1999):

$$Asset\ return = R_i = \frac{P_i^1 + D_i - P_i^0}{P_i^0}$$

$$Portfolio\ return = R_p = \sum_{i=1}^{n} w_i R_i$$

Where,

P_i^t = closing price of an asset at time t

D_i = Dividends of an asset

w_i = weight of an asset within a portfolio

The formulas for risk are given by (Bodie et al., 1999):

$$Asset\ risk = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}\left[R_{ij} - \mu)\right]^2}{n-1}}$$

$$Portfolio\ risk = \sigma_p = \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2Cov(R_1, R_2)}$$

$$= \sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\rho_{1,2}\sigma_1\sigma_2}$$

Where,

μ (=E(R)) = average of returns

n = number of returns considered

Cov = covariance

ρ = correlation coefficient

The formula for correlation coefficient (Bodie et al., 1999):

$$\rho = \frac{Cov(r_1, r_2)}{\sigma_1\sigma_2}$$

$$= \frac{\sum_{i=1}^{n}[(R_1 - E(R_1))(R_2 - E(R_2))]}{\sigma_1\sigma_2}$$

We then constructed an efficient frontier by plotting the return against risk of each portfolio. As we can see in the Figure 3.4.2 below, the efficient frontier lies above the global minimum-variance portfolio, portfolios below that are inefficient.

Then, various calculations were conducted to achieve a capital allocation line (CAL) that is tangent to the efficient frontier. The formula for CAL (Bodie et al., 1999):

$$CAL: R_p = R_f + S_p\sqrt{\sigma_p}$$

Where,

R_f = return of a risk free asset

S_p = Sharpe ratio

The point that touches the efficient frontier is the optimal portfolio that maximizes the Sharpe ratio, as we can see in the Figure 3.4.2. below. The Sharpe ratio, developed by Sharpe (1994), is a "reward-to-variability ratio" that helps investors understand the return of a portfolio compared to its risk. The formula of Sharpe ratio (or the slope of CAL) is given by:

$$S_p = \frac{R_p - R_f}{\sqrt{\sigma_p}}$$

Assuming the return of a risk free asset is equal to 0, in Europe this is the current case, this corresponds to:

$$S_p = \frac{R_p}{\sqrt{\sigma_p}}$$

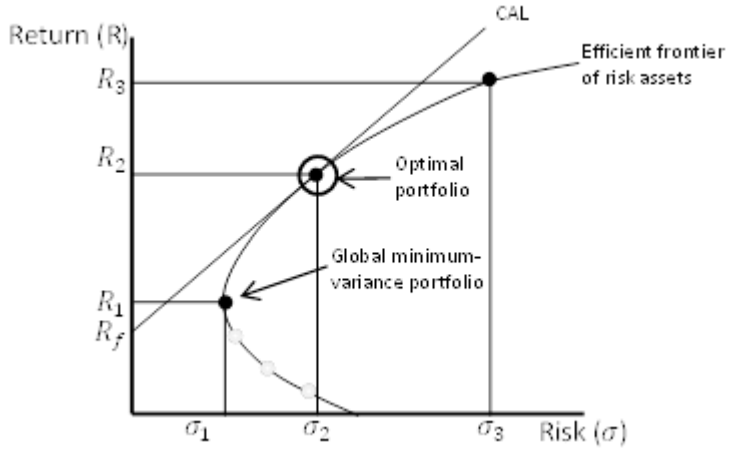

Figure 3.4.2 – CAL Tangent to Efficient Frontier

3.4.3. Investment Strategy

The investment strategy considered for this project is summarized in Table 3.4.3. In this project the main objective is to compare both models' portfolios to understand the impact that predictions could have in their performance, so we wanted to keep the investment strategy as simple as possible therefore no trading costs were considered. Within each period (15 in total), one portfolio was constructed by each model (30 in total), classical model and modified model, and these portfolios are independent of one another so that the performance is analyzed by period.

Another assumption is that volatility is considered as a measure of risk. For the classical model, the volatility is based on historical prices but for the modified model, the volatility considered is partially estimated as it accounts for one estimated price and two historical prices. This was due to only one price being estimated per period and to consider the near full impact of the estimated prices in the construction of the portfolios.

<table>
<tr><th>Model</th><th>Optimization Goal</th><th>Assumptions</th><th>Inputs</th></tr>
<tr><td>Classical Markowitz model</td><td rowspan="2">Maximization of the Sharpe ratio</td><td rowspan="2">- No costs considered
- Portfolios in each period are independent
- Volatility considered as measure of risk</td><td>Expected return and risk are obtained from historical prices</td></tr>
<tr><td>Extended version of Markowitz model</td><td>Expected return and partially estimated risk are obtained from the RF model</td></tr>
</table>

Table 3.4.3 – Project's Investment Strategy

4. RESULTS

4.1.1. Accessing RF Model Reliability

To test the accuracy of the predictions of the RF algorithm, the following common metrics were used (Hyndman and Athanasopoulos, 2018).

- **Mean absolute error**

Mean absolute error (MAE) measures the average over the data sample of the absolute differences between the predictions and the actual observations. It is given by:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Where,

y_j = is the true value
$\hat{y}_j$ = is the prediction

- **Mean absolute percentage error**

Mean absolute percentage error (MAPE) measures the percentage average of the absolute differences between predictions and the actual observations. It is given by:

$$MAPE = \frac{100}{n}\sum_{j=1}^{n}\left|\frac{y_j - \hat{y}_j}{y_j}\right|$$

- **Accuracy**

The accuracy of a prediction can be extracted from the MAPE:

$$Accuracy = 1 - MAPE$$

The Table 4.1.1 shows the average accuracy of the predictions of the 15 periods by company. The RF model predictions were very reliable for some stocks namely Altria (MO) with 98.3% average accuracy and not very reliable for other stocks namely United Health (UNH) with 78.2% average accuracy. On average the RF model predictions were to some extent reliable with 88.9% accuracy. As mentioned in the literature review, various studies attested that the RF model predictions have the best average performance overall (Weng et al., 2018) and it was demonstrated in this paper that its performance was quite high, almost 90%.

	MO	AXP	CTXS	DIS	ECL	FDX	F	OXY	UNH	XEL	Avg
MAE	0.7	14.4	17.7	6.1	29.6	22.0	1.0	1.6	53.5	4.4	15.1
MAPE (%)	1.7	13.8	17.2	4.7	18.3	10.2	11.2	2.9	21.8	9.0	11.1
Accuracy (%)	98.3	86.2	82.8	95.3	81.7	89.8	88.8	97.1	78.2	91.0	88.9

Table 4.1.1 – Average Accuracy of Predictions

4.1.2. Performance Measures

The return, risk and Sharpe ratio figures were calculated again using actual prices as at forecast date for each period. The performances of the portfolios of each algorithm were analyzed using the following measures:

- Comparing the return and risk of each portfolio

Portfolios' return and risk were calculated as explained above as at forecast date. As we can see in the graphs below, the extended version of Markowitz model surpassed the Markowitz model in terms of risk and return. For 12 out of 15 periods the modified model had higher returns and in terms of risk, for 8 out of 15 periods it had lower risk.

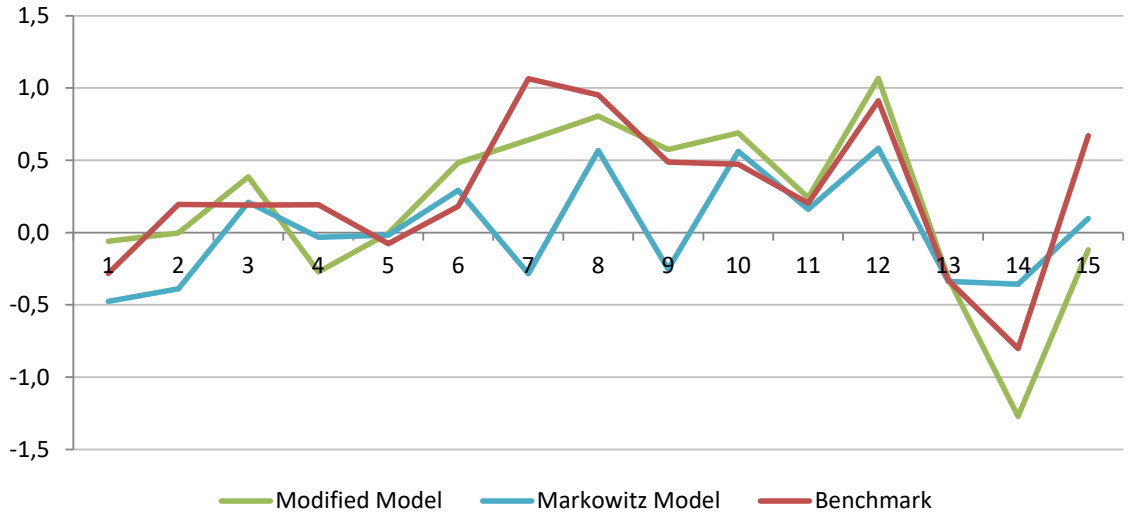

Figure 4.1.2 – Portoflio Return (%)

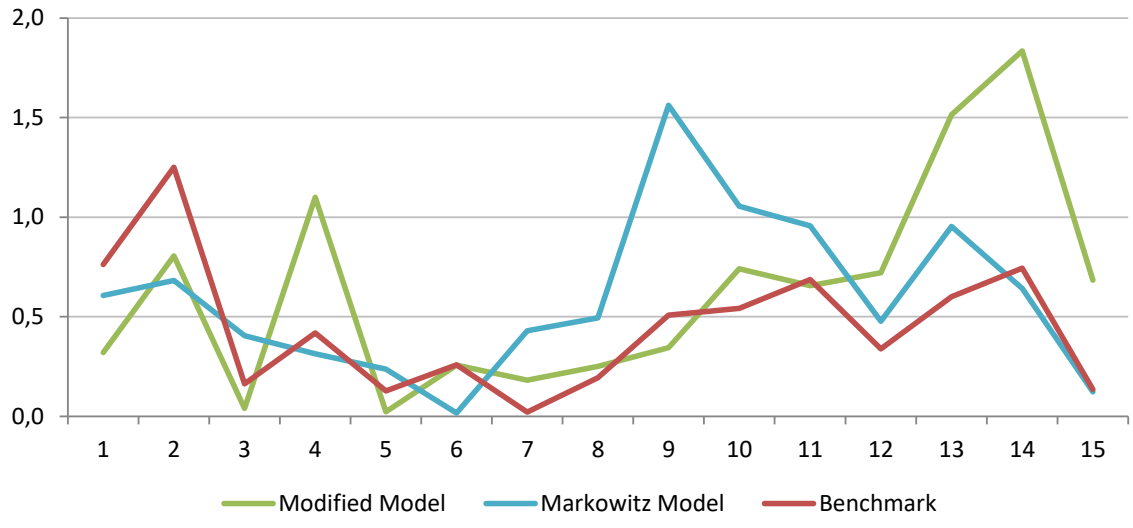

Figure 4.1.2 – Portoflio Risk (%)

- Comparing the Sharpe ratio of each portfolio

Sharpe ratio is used by many investors to evaluate the performance of their portfolios since it measures the risk adjusted to return. Portfolios' Sharpe ratio was calculated as explained above as at forecast date. As we can see in the graph below, the modified model continued to surpass the Markowitz model in terms of Sharpe ratio. For 10 out of 15 periods the modified model had higher Sharpe ratio.

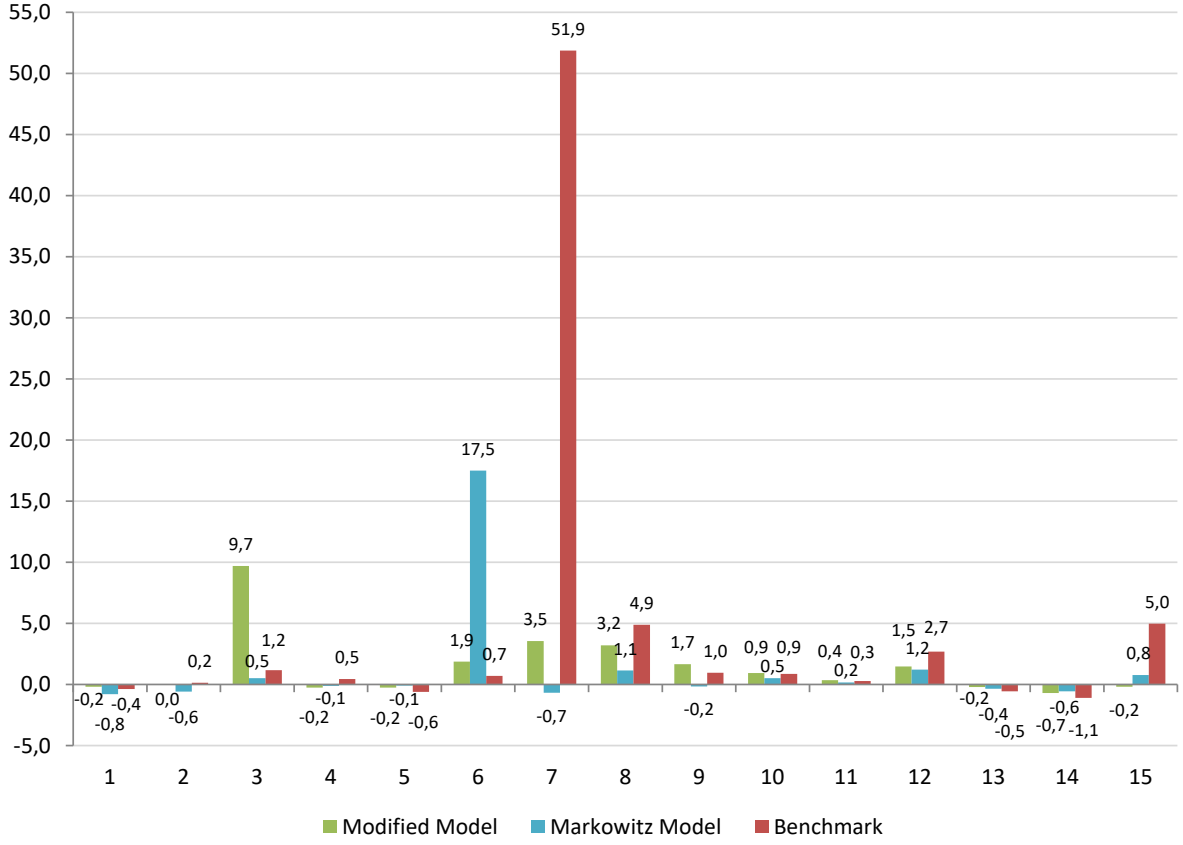

Figure 4.1.2 – Portoflio Sharpe Ratio

- Comparing the portfolios of each model to a benchmark

Benchmarks are used by investors to analyze how their portfolios' are performing compared to other market segments. The benchmark used in this project is the S&P 500 since the companies selected are from this index. As we can see from the graphs above, the modified model outperformed the Markowitz model and the benchmark in terms of return for 9 out of 15 periods. The modified model performed in line with the benchmark as the Sharpe ratio was highest for both 6 out of 15 periods each. In terms of risk, both models and the benchmark had 5 out 15 periods each where the risk was lowest.

The results in this paper are in line with studies that have been conducted namely in Tadlaoui (2018) where the predictions had a very positive impact on the portfolios which lead to outperformance of the RF model, among other studies.

5. DISCUSSION AND CONCLUSIONS

Many researchers believe that it is impossible to outperform the market by trying to predict future stock prices because it is very volatile and correlated with real time events. This project aims to demonstrate a feasible way to use machine learning models as assistance to human decision-making processes. The results validate that the predictions of the RF model are not very reliable for some stocks but on average the model was 88.9% accurate, which in general is quite high.

When the performance of the portfolios of both models were compared, it was clear that the modified model outperformed the classical model in terms of risk, return and Sharpe ratio for most periods. When compared to the benchmark, the modified model also outperformed in terms of return but for risk and Sharpe ratio it performed in line with the benchmark.

The results support the idea that ML can in fact improve portfolio performances. The impact that the predictions had on the portfolios gives great incentive to further develop the model to improve accuracy by adding, for instance, more technical indicators (Xinjie, 2014), input external data that influences stock prices (Li et al., 2014) or even changing the model's parameters. One could change the number of trees in the forest (for this project 100 trees were used as it was the model's default) changing the tree weights by increasing the ones with better performance (Winham et al. (2013) and Bernard et al. (2012)), among other changes.

ML models are very flexible and can adapt to various problems just by adjusting the criteria and parameters. They are also highly efficient when doing repetitive tasks and are able to identify patterns that may not be obvious to humans; models can extract information from unstructured data sources; and unlike statistical models, ML models are able to improve themselves by readjusting according to the data (Bartram et al. (2020)). This gives great incentive to extend the use of ML models in portfolio construction and monitoring. If companies included ML in their trainings to employees this could complement the methods already used to improve results.

The study's main limitation was the accuracy of the predictions, in each period for some companies the accuracy was quite high but it was not consistent for all companies. As discussed above, for a prediction to have higher accuracy, some alterations can be done to the model and one fairly easy change could be to increase the number of trees. This was attempted but unfortunately the model became very slow as more trees were added and no results could be extracted therefore only 100 trees were used for each prediction, which was the model's default.

Another study's limitation was the fact that only one stock price was predicted per period for each company which led to only one estimated price being used to calculate expected volatility and the remaining two prices used were historical. This lead to what was derived from the RF model and inputted in the Markowitz model was not be fully estimated so we could not fully analyze the predictions' impact on the classical model. The prediction of more prices was also attempted but as more prices were estimated, the accuracy ended up decreasing which was not ideal therefore, to simplify, only one price was estimated per period for each company.

One final limitation was the over simplification of the investment strategy. In the real world, the investment strategy has to consider trading costs, investor's risk aversion and objectives, and portfolio management. This project was simplified to the point that all these factors were ignored so

that the only thing being considered was the selection of portfolios based on one optimization goal. This was done so that the main focus of the project was the impact estimated prices had on the performance of the portfolios.

Despite these limitations, the modified model was still able to outperform the classical Markowitz model. Therefore this project demonstrates that ML models should be incorporated in the classical portfolio optimization models to obtain better results. Unfortunately due to the simplification of the project, private investors and experts are not able to use this model as it excluded some factors that are important to be considered in the real world, namely trading costs. In future research, these factors should be included to make the model as close to reality as possible. It would also be interesting for future research to add some inputs to the modified model, as discussed above, that influence stock prices to increase the accuracy of predictions and to be able to increase the number of estimated stock prices. Instead of only predicting one stock price, as done in this project, at least 3 stock prices should be predicted to calculate the expected volatility with only estimated prices and to evaluate the full impact estimated prices have in the construction and performance of the portfolios.

6. REFERENCES

Alpaydın, E. (c2010). Introduction to Machine Learning. (2nd ed.). England: The MIT Press.

Appel, G., & Dobson E. (2008). Understanding MACD (Moving Average Convergence Divergence). (1st ed.). USA: Traders Press Inc.

Bartram, M. S., Branke, J., & Motahari, M. (c2020). Artificial Intelligence in Asset Management. USA: CFA Institute Research Foundation.

Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, R. S. (2019). Predicting the direction of stock market prices using tree-based classifiers. The North-American Journal of Economics and Finance, 47, 552–567.

Bernard, S., Adam, S., & Heutte, L. (2012). Dynamic Random Forests. Pattern Recognition Letters, Elsevier, 33 (12), 1580-1586.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25(2), 197–227.

Bodie, Kane & Marcus. (1999). Investments. (5th ed.). USA: McGraw−Hill Primis.

Breiman, L. (1996). Bagging Predictors. Machine Learning, 24(2), 123-140.

Breiman, L. (2001). Random Forests. University of California, Berkeley, CA.

Breiman, L., Friedman, H. J., Olshen, A. R., & Stone, J. C. (1984). Classification and regression trees. USA: Brooks/Cole Publishing.

Dai, Y. & Zhang, Y. (2013). Machine Learning in Stock Price Trend Forecasting. Stanford University, Stanford, USA.

Fama, F. E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(1), 383-417.

Geurts, P., & Louppe, G. (2011). Learning to rank with extremely randomized tree. JMLR: Workshop and Conference Proceedings, 14, 4961.

Granville, E. J. (1976). Granville's new strategy of daily stock market timing for maximum profit. USA: Prentice-Hall.

Hyndman, J. R. & Athanasopoulos, G. (c2018). Forecasting: Principles and Practice. (2nd ed.). Australia: OTexts.

Ishwaran, H., Kogalur, B. U., Blackstone, H. E. & Lauer, S. M, (2008). Random Survival Forests. The Annals of Applied Statistics, 2(3), 841–860.

Khaidem, L., Saha, S., & Dey, R. S. (2016). Predicting the direction of stock market prices using random forest. Applied Mathematical Finance, 1(5), 1-20.

Lane, C. G. (1984). Lane's Stochastics. Technical Analysis of Stocks and Commodities magazine. 87-90.

Li, H., Yang, Z., & Li, T. (2014). Algorithmic Trading Strategy Based On Massive Data Mining. Stanford University, Stanford, USA.

Malkiel, G. B. (2003). The efficient market hypothesis and its critics. The Journal of Economic Perspectives, 17(1), 59-82.

Markowitz, H. (1952). Portfolio Selection. The Journal of Finance, 7(1), 77-91.

Marsland, S. (c2015). Machine Learning: An Algorithmic Perspective. (2nd ed.). England: CRC Press.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12(85), 2825−2830.

Prado, L. M. (2018). Advances in Financial Machine Learning. USA: John Wiley & Sons, Inc.

Ravinder, H. V. (2013). Determining the Optimal Values of Exponential Smoothing Constants – Does. Solver Really Work?. American Journal Of Business Education, 6(3).

Sharpe, F. W. (1994). The Sharpe Ratio. The Journal of Portfolio Management, 21(1), 49-58.

Tadlaoui, G. (2018). Intelligent Portfolio Construction: Machine-Learning enabled Mean-Variance Optimization. Imperial College London, London, England.

Weng B., Lu L., Wang X., Megahed F., & Martinez, G. W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. Expert Systems With Applications. 112, 258-273.

Wilder, W. J. (1978). New concepts in technical trading systems. USA: Trends Research.

Winham, J. S., Freimuth, R. R., & Biernacka, M. J. (2013). A Weighted Random Forests Approach to Improve Predictive Performance. Stat Anal Data Min. 6(6), 496–505.

Xinjie, D. (2014). Stock Trend Prediction With Technical Indicators using SVM. Stanford University, Stanford, USA.

Yahoo finance. (c2020). Yahoo Finance. Retrieved 15 June, 2020, from https://finance.yahoo.com/

Zhang, X., Li, A., & Pan, R. (2016). Stock trend prediction based on a new status box method and adaboost probabilistic support vector machine. Applied Soft Computing, 49, 385–398.

Zhang, Y., Li, X. & Guo, S. (2018). Portfolio selection problems with Markowitz's mean–variance framework: a review of literature. Fuzzy Optimization and Decision Making, 17(2).