Tilburg University

PREDICTING STOCK PRICE MOVEMENTS BASED ON INTERNAL FINANCIAL AND MACROECONOMIC INDICATORS

BASTIAAN MONNEE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2047093

COMMITTEE

dr. Richard Starmans
dr. Bruno Nicenboim

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 10, 2023

ACKNOWLEDGMENTS

I want to thank my supervisor, dr. Richard Starmans for guiding me through the research process. The weekly Zoom meetings were beneficial for clarifying things, getting confirmation, and sparring to get new insights. Furthermore, I would like to thank dr. Bruno Nicenboim for taking the role of the second reader. Finally, I would like to thank the people involved in the databases, open-source software, and packages used for this research for making this freely available.

PREDICTING STOCK PRICE MOVEMENTS BASED ON INTERNAL FINANCIAL AND MACROECONOMIC INDICATORS

BASTIAAN MONNEE

Words: 8328

Abstract

This research examines the extent to which internal financial and macroeconomic indicators can correctly classify if the stock price goes up or down in the following year. This research has been applied to design a trading strategy for individuals with less knowledge about the market, as the data used for this study is publicly available. The research builds upon previous research into the effect of internal financial and macroeconomic indicators on stock price movements. However, the effect of both internal financial and macroeconomic indicators together and making a binary prediction problem for this through machine learning has rarely been put into practice. Seven machine learning algorithms (Decision Tree, Random Forest, Gradient Boosting Machine, Support Vector Machine, k-Nearest Neighbor, Naïve Bayes and Logistic Regression) are trained on 43 features and 19,809 observations of US companies between 2014 and 2018. This research distinguishes itself even more from previous studies due to the extensive dataset and algorithms performed. All models used in this study outperform the baseline model, the majority class classifier. In addition, the performances of the different models are close to each other. The best-performing model, Random Forest, has an accuracy of 70.24%, which is 16.09% higher than the baseline model. Features such as the Mortgage rate, Earnings per Share, Return on Equity, and Net income play an important role. In conclusion, internal financial and macroeconomic indicators play a reasonably large role in the binary stock price movement the following year.

1 INTRODUCTION

This chapter starts with discussing the problem statement, in which both the societal and practical relevance are discussed. In addition, the research questions will be briefly discussed. Finally, the main findings will be formulated.

1.1 *Problem statement*

There is a wide range of research on predicting stock prices, such as predicting stock prices the next day based on price developments (Henrique, Sobreiro, & Kimura, 2018). In addition, there is much research in which macroeconomic indicators are considered (Abbas, Hammoudeh, Shahzad, Wang, & Wei, 2019; Ma, Lu, Liu, & Huang, 2022). There is much research on stocks since many people and companies invest in stocks. However, they want to prevent the stock from decreasing in value. This research attempts to explore a side that is less known, predicting stock price movements, whether the stock prices go up or down in the next year, by internal financial indicators over the past year but also considering the macroeconomic indicators. These are indicators that are expected to influence stock prices as internal financial indicators provide insight into the state of a company's economy. These internal financial indicators include, for example, a company's revenue and market cap, but also more ratio-related indicators such as the Earnings per Share (EPS) and Debt/Equity-ratio. While the macroeconomic indicators provide insight into the state of a country's economy, including, for example, inflation, the housing market, and unemployment. In general, there is little research on the predicting problem of internal financial and macroeconomic indicators together on stock prices in the next year. This study differs from previous studies since it is a binary classification problem applied for an entire year, and no exact price movement is predicted within a short time range. In addition, the effect investigated in this research has only been investigated in a few countries, but few prediction models have been made. Furthermore, previous studies used smaller datasets. While, this research dataset contains more than 200 internal financial and macroeconomic indicators. Finally, seven algorithms are used in this research.

The stock price movement is an interesting problem to gain more insight into since investors generally decide their investments rationally. Making rational decisions requires much information about companies, which takes time to collect. Trading companies are specialized in this and therefore have much knowledge about the market. However, individuals can also make better decisions with internal financial and macroeconomic indicators because this data is publicly available. Besides being publicly available, the data is easy to generate for individuals. Macroeconomic indicators are published annually by the U.S. Bureau of Economic Analysis, and the internal financial indicators are published yearly by the 10-K fillings (Carbone, 2019).

The societal relevance of this research revolves around the fact if internal financial and macroeconomic indicators can be used to predict whether a stock will rise or fall in price the following year for individual people to apply this in their trading strategy. Individuals can better distinguish successful and unsuccessful stocks using the proposed model of this research. Besides the fact that investors can earn profit by applying this model in their trading strategy, this research is also relevant from a higher moral principle of doing business fairly and transparently.

1.2 *Research strategy*

The following question is central to this research:

> *To what extent is it possible to predict stock price movements in the next year based on last year's internal financial and macroeconomic indicators?*

In order to build a prediction model for stock price movements, it is essential to see which internal financial and macroeconomic indicators affect stock price movements and which algorithms can be used for this prediction problem. Therefore, an answer to this central research question will be obtained by answering the following sub-questions and ultimately applying the answers in the model:

RQ1 *Which internal financial indicators are important in the binary prediction of stock price movements?*

RQ2 *Which macroeconomic indicators are important in the binary prediction of stock price movements?*

RQ3 *Which machine learning algorithm can best predict binary stock price movements?*

Firstly, to answer these research questions, the individual effect of internal financial and macroeconomic indicators on stock price movements in previous research will be discussed. Secondly, the effect of internal financial and macroeconomic indicators together on stock price movements in previous research will be discussed. Thirdly, there will be a literature section about which algorithms have been used in previous studies for a classification problem such as this one within the stock market. Finally, all three sub-questions will also be tested in practice, with the literature in mind, to arrive at the best prediction model and be able to answer the main research question and compare the answers with the previous literature. This will be done by model comparison of the seven algorithms based on different evaluation metrics, where the various errors will be analyzed using the confusion matrices. Since this research uses an extensive dataset, wrapper methods are used for the feature selection process.

1.3 *Main findings*

The main finding of this research is that internal financial and macroeconomic indicators can be used in predicting annual stock price movements using machine learning algorithms. In this study, macroeconomic indicators, especially the Mortgage rate, play the most prominent role. Several internal financial indicators also play a role in predicting the stock price movement. For example, the Debt/Equity-ratio, Earnings per Share, Return on Equity, Asset turnover ratio, Net income, Prices to sales ratio, Cash ratio, and Revenue growth. Within this research, seven machine learning algorithms are examined. The Random Forest algorithm best predicts whether the stock price will go up or down next year.

2 LITERATURE REVIEW

This chapter will start with a section about the stock market. Secondly, the individual effect of internal financial and macroeconomic indicators on stock prices will be discussed. After these sections, the effect of the internal financial and macroeconomic indicators together on stock prices will be discussed. Furthermore, the stock price prediction methods in previous studies will be discussed. Finally, the differences with related work will be discussed.

2.1 *Stock market*

Stocks of listed companies are traded on the stock market. The main reason for a company to be listed on the stock market is to raise capital and name recognition (Times, 2022). The main reason for investing in stocks is to make profit. In addition, Aspara and Tikkanen (2011) concluded that to financial expectations, affect-based feelings also play a role in investing in stocks. Both internal and external factors play a role in stock price movements, which will be discussed in the following sections.

2.2 *Effect of internal financial indicators on stock prices*

First, the effect of internal financial indicators on stock price movements will be discussed. Internal financial indicators give insight into the state of a company's economy. Ali, Mubeen, Lal, and Hussain (2018) concluded that Logistic Regression (LR) can be used to predict stock price performance using the Book-to-price ratio, Current Ratio (CR), Debt/Equity-ratio, Earnings per Share, Sales growth, and Return on Equity (ROE) of companies. While Arkan (2016) also concluded that the stock price could be predicted by some financial ratios, such as the Return on Assets (ROA), ROE, and EPS. According to Arkan (2016), the important financial ratios differ per sector. In addition, Vedd and Yassinski (2015) found significant effects of the Assets turnover ratio, Debt ratio, and Firm size on the stock prices, while there were no significant effects for the Net profit margin (NPM), ROE, CR and Cash flow from operations. In contrast, Öztürk and Karabulut (2018) concluded that the NPM and P/E-ratio significantly positively affects stock prices, while the CR has no significant effect. In line with Vedd and Yassinski (2015), Jiang, Wang, Li, Wang, and Huang (2019) concluded that the Total asset turnover positively affects stock prices. Jiang et al. (2019) also finds different movements between the different sectors. Banchuenvijit (2016) concluded that the CR, NPM and Total assets turnover ratio positively affects stock prices, while the Debt/Equity-ratio negatively affects stock prices. Puspitaningtyas (2017) investigated the effect of financial performance on stock prices. Partly contradictory to the previous results, only EPS significantly affects stock prices, while Sales growth indicators, CR, and ROE have no significant effect. Also, in contract with most research, Yuliarti and Diyani (2018) concluded that the Market book ratio positively affects stock prices, while Cash flow from financing activities negatively affects stock prices. Within the research, Firm size, ROE, CR, Cash flow from operating and investing activities seems to have no significant effect on stock prices.

To conclude, the literature shows the Debt/Equity-ratio (Ali et al., 2018; Banchuenvijit, 2016), EPS (Ali et al., 2018; Arkan, 2016; Puspitaningtyas, 2017), Total asset turnover ratio (Banchuenvijit, 2016; Jiang et al., 2019; Vedd & Yassinski, 2015) significantly affects stock price movements. However, the theory also shows contradictory results. According to Ali et al. (2018); Banchuenvijit (2016), the CR significantly affects stock price movements, while this is not the case according to Öztürk and Karabulut (2018); Puspitaningtyas (2017); Vedd and Yassinski (2015); Yuliarti and Diyani (2018). The same applies to the NPM, Banchuenvijit (2016); Öztürk and Karabulut (2018) found significant effects, while Vedd and Yassinski (2015) found no significant effect. Finally, Ali et al. (2018); Arkan (2016) concluded that the

ROE has a significant effect, while according to Puspitaningtyas (2017); Yuliarti and Diyani (2018), the ROE has no significant effect.

2.3 *Effect of macroeconomic indicators on stock prices*

Secondly, the effects of macroeconomic indicators on stock price movements will be discussed. Macroeconomic indicators give insight into the state of a country's economy. Abbas et al. (2019) found strong interactions between returns and volatilities and macroeconomic indicators, such as Exchange rates, Industrial production, Inflation, Interest rates, Money supply, and Oil prices. In line with Abbas et al. (2019), Ma et al. (2022) concluded that stock market returns can be predicted using macroeconomic attention indices (MAI). These MAI were mainly created by Fisher, Martineau, and Sheng (2022) based on the following macroeconomic indicators: Credit ratings, The housing market, Inflation, Monetary policy, Oil, Output growth, Unemployment, and the US dollar. Besides, Chen (2009) concluded that it helps to predict recessions in the stock market using macroeconomic indicators. A strong effect was found, especially in Inflation rates and Yield curve spreads. Furthermore, Fromentin, Lorraine, Ariane, and Alshammari (2022) found asymmetric bidirectional causality between macroeconomic indicators and the stock market. Fromentin et al. (2022) concluded that this effect is more prevalent during recessions.

To summarize, all studies find significant effects of macroeconomic features on stock price movements. For example, studies show inflation (Abbas et al., 2019; Ma et al., 2022), exchange rates (Abbas et al., 2019; Chen, 2009; Ma et al., 2022) and unemployment (Ma et al., 2022) as features that influence stock price movements.

2.4 *Effect of both internal financial and macroeconomic indicators on stock prices*

Finally, a few times, research has been done into the effect of internal financial and macroeconomic indicators together on stock price movements. However, this has been done with fewer features and different models or even without a prediction model. Kwag and Kim (2013) investigated the effect of financial ratios on stock prices, controlling for macroeconomic indicators through LR. They concluded that financial ratios influence stock price movements, overcoming macroeconomic indicators' effects. The accuracy score in this research is 58.30%. In addition, Kwag and Kim (2013) concluded that there are time and industry effects. Karakus and Bozkurt (2017) found a negative effect of the Debt-ratio on stock returns and a positive effect of the ROA and Net working capital turnovers. Regarding the macroeconomic indicators, a positive effect was found for

Unemployment, Gross domestic product, Net inflows of portfolio equity, and Exchange rates. While Inflation negatively affects stock returns. Ulandari and Damayanthi (2021) found some other effects. In their research, the Exchange rate, Interest rates, and Profitability positively affect stock returns, while Inflation, Leverage, and Liquidity do not affect stock returns.

2.5 *Stock price prediction methods*

This section will discuss which algorithms have been used in other studies for stock price movement prediction. Firstly, supervised learning is used in this study since the outcome variable is known. Secondly, the exact stock price movement is predicted in many studies (Henrique et al., 2018), so through a regression problem. In contrast, this research is a binary classification problem of whether the stock price is going up or down. Therefore, only classification algorithms will be considered. For predicting stock price movements, many different classifiers have been tested in previous studies. Research shows that deciding which algorithm works best depends on the evaluation criteria (Dash, Samal, Rautray, & Dash, 2019). The section below will discuss the algorithms of previous studies into stock price movement predictions.

Ravikumar and Saraf (2020) examined whether the stock price moves up or down the next day using six classification algorithms, namely, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), LR, Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF). LR gave the best results, with an accuracy of 68.41%. In addition, the SVM and NB algorithms were close to the results of the LR algorithm. In line with this, Ali et al. (2018) concluded that LR could classify stock price performance using internal financial indicators. However, Zhong and Enke (2017) concluded that ANN works better than LR in predicting stock price movements. Nevertheless, since Ali et al. (2018); Ravikumar and Saraf (2020) show that LR is a good algorithm for this prediction problem, LR will be examined.

In addition, NB and DT are two classification algorithms that have been compared by Hutapea, Samuel, and Sitorus (2019). Their research predicted stock price movements, and the DT algorithm appears to perform better than the NB algorithm. In contrast, Ravikumar and Saraf (2020) concluded that the NB algorithm had good results, with an accuracy of 67.10%. Because of this, NB will be tested within this research to see the results.

Besides the DT algorithm, the RF algorithm is suitable for a classification problem. Basak, Kar, Saha, Khaidem, and Dey (2019) concluded that the RF algorithm is excellent for predicting stock price movements. While Illa, Parvathala, and Sharma (2022) concluded that both the RF and SVM are good algorithms for predicting the stock market. The RF algorithm performs best of the two in their research. Since, according to Basak et al. (2019); Illa et al. (2022), the RF algorithm is a good algorithm for predicting stock price movements, it will be tested in this research. Since RF is less interpretable than the DT algorithm and the DT algorithm performs well in previous research (Hutapea et al., 2019), the DT algorithm is also tested within this research.

In addition, the SVM algorithm has been used in various studies. Hu, Zhu, and Tse (2013) concluded that the SVM algorithm is an excellent algorithm for predicting stock prices. Z. Li and Tam (2017) tried to predict stock price movements and concluded against expectations that the SVM algorithm outperformed recurring neural networks at low volatile stocks. The recurring neural networks did outperform SVM at highly volatile stocks. In addition, Heo and Yang (2016) concluded that stock prices can be accurately predicted using SVM and financial information and the predictability decreases over time. In line with these results, Reddy (2018) used the SVM algorithm to predict stock price movements. One disadvantage of the SVM algorithm is the lack of transparency in the results (Karamizadeh, Abdullah, Halimi, Shayan, & Mohammad, 2014). However, since several studies show the SVM algorithm is good for classifying stock price movements, it will be used in this study.

The KNN algorithm can also predict stock price movements (Alkhatib, Najadat, Hmeidi, & Shatnawi, 2013). Subha and Nambi (2012) conducted research into predicting stock index movements. It was concluded that the KNN algorithm performs well, even outperforming LR. In general, KNN works poorly for large datasets. However, since KNN has proven to be a good algorithm for predicting stock price movements, it will be tested in this study.

Basak et al. (2019) concluded that the Gradient Boosting classifier (GBM) is excellent for predicting stock price movements. A disadvantage of GBM is that the results are less interpretive. However, since other algorithms, such as LR, are more interpretive, this algorithm is being tested in this research.

Lastly, it is interesting that Kumar, Dogra, Utreja, and Yadav (2018) investigated the prediction of stock market trends. Five models were implemented in this study, namely KNN, NB, RF, SoftMax, and SVM. The research shows that the NB algorithm is best for small datasets, and the RF algorithm is best for large data sets. The research also shows that reducing the number of technical indicators reduces the accuracy of all algorithms.

To conclude, the algorithms DT, GBM, KNN, LR, NB, SVM, and RF will be tested within this study, as they have been shown in previous studies to be relevant in predicting stock prices and are suitable for a supervised classification problem. Not all algorithms are best for interpretation purposes, but they are still used as some other algorithms, such as LR and DT, suit well for interpretation purposes.

2.6 *Differences with related work*

This research predicts whether the stock price will go up or down in the following year based on last year's internal financial and macroeconomic indicators. As discussed, there is much literature about stock prices. For example, several studies have investigated the effect of internal financial indicators on stock price movements (Ali et al., 2018; Arkan, 2016), but also the effect of macroeconomic indicators on stock price movements (Abbas et al., 2019; Ma et al., 2022). In addition, there is research into the effect of internal financial and macroeconomic indicators together on stock price movements (Karakus & Bozkurt, 2017; Kwag & Kim, 2013).

This study differs for several reasons from the other studies. Firstly, this research is a binary classification problem. So, this research aims to predict stock price movements over a year instead of more frequent window and exact price movements. This is the difference between short and long-term investing within the stock market. There is much research on predicting daily stock price movements (Henrique et al., 2018). However, there is less research on predicting binary stock price movements over an entire year. The study takes this approach, as this model also applies to investors with less background information than, for example, investment companies. This is the case since the data used in this study is published publicly annually. Secondly, there is much research on the effects of internal financial and macroeconomic indicators, but the previous research is limited on prediction algorithms that use both internal financial and macroeconomic indicators. Thirdly, the dataset used for this research is extensive, with over 200 features. Previous studies did not use such a large dataset, meaning other effects may emerge in this study. Lastly, seven algorithms are used in this research. In previous studies, the algorithms were more limited.

To conclude, this study differs from previous studies since it is a binary classification problem over a year. In addition, many (different) algorithms are used, the dataset has many features, and both macroeconomic and internal financial indicators are used to predict stock price movements.

3 METHODOLOGY AND EXPERIMENTAL SETUP

This chapter will start by explaining the datasets and software used. Hereafter, the data preprocessing, feature selection process, training process of the algorithms, and evaluation metrics will be discussed. Finally, this chapter is concluded by visualizing the research through a flowchart.

3.1 *Initial dataset*

Two datasets will be mainly used in this research. One dataset revolves around the internal financial indicators of US companies from Kaggle (Carbone, 2019). There is a separate dataset for each year, and the datasets are available for the years 2014 to 2018. These five datasets were merged into one large dataset. Within this dataset, there are 222 publicly available US companies' internal financial features, such as EPS, Revenue, and Asset growth. In addition, the dataset's target variable is whether the company's stock price has gone up (1) or down (0) the following year. In total, the dataset consists of 22,077 observations.

Furthermore, there is a second dataset for the macroeconomic indicators from Kaggle (Mirashi, 2022). This dataset consists of seven monthly macroeconomic indicators (CPI, Mortgage rate, Unemployment rate, NASDAQ, Disposable income, Personal consumption expenditure, and Personal savings) in the US from 1980 to 2022.

3.2 *Software*

The programming language R is used in this research. Several packages were used within the R program. Table 4 in Appendix B (page 34) shows which packages were used and for what purpose.

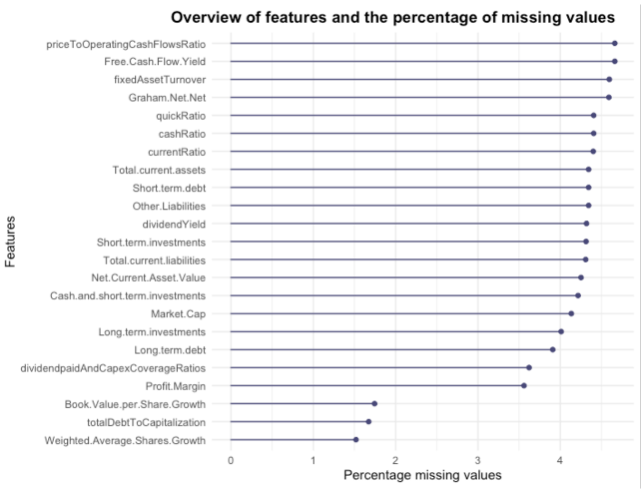
3.3 *Data preprocessing*

In order to arrive at a good dataset, the dataset has been preprocessed. In the macroeconomic dataset, the yearly growth rate of all features is calculated. These calculated growth rates will ultimately be used as features. For the internal financial indicators dataset, only feature selection has been applied, while no features are calculated based on other features, for example. There was only one categorical feature, Sector, containing eleven different sectors. This variable has been converted into dummy variables. After these steps, the two datasets were merged by year.

After these steps, the outliers were treated. It was decided to apply the winsorization method. In the winsorization method, extreme values are replaced by the nearest non-extreme values (Wilcox, 2005). This study applied the winsorization method at the 1% level. Applying the winsorization method can lead to bias, as the impact of extreme values can be undervalued. However, this method was chosen because outliers can have too much influence and therefore lower the evaluation metrics of models and increase the training time. In addition, this method suits this research because of the large dataset compared to other methods of dealing with outliers.

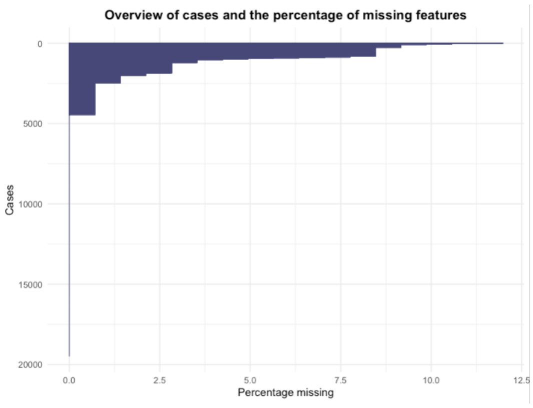
Furthermore, it was checked whether any feature had more than 15% missing cases. When this was the case, the features were removed. Moreover, when an observation has more than 10% of features missing features, this observation is dropped. Hereafter, the correlations between the different features were examined. This showed that several features were the same but with different names. In each case, the one with the most missing values was dropped. Besides, the features with a very high correlation (>0.975) were examined. One of the two was removed based on which had the fewest missing cases or was found to be important in previous studies. After these steps, there were still features with missing values. Figure 1 below shows the features that have more than 1% of observations with missing values.

Figure 1: Overview of features and the percentage of missing values.



In addition, Figure 2 below shows the observations and their missing feature rates. It is visible that there are no observations with more than 10% missing features. Besides, the figure shows that most observations have no missing features. To deal with the remaining missing values, the missing values were replaced using multiple imputation. Multiple imputation has been used to get plausible imputations of the missing values, while uncertainty is accurately reflected (Freedman & Wolf, 1995; Schafer, 1998). In addition, compared to other methods, with this method, much data remains available.

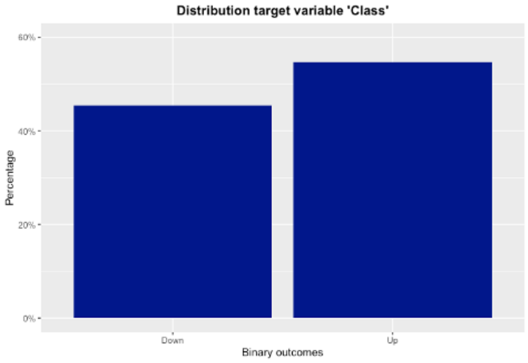
Figure 2: Overview of cases and the percentage of missing features.



Furthermore, multicollinearity is dealt with since many features are based on each other in the dataset. Multicollinearity has been dealt with as this research also aims to see which features influence the prediction of stock price movements, and for instance, LR assumes minimal multicollinearity (Stoltzfus, 2011). Features with a Variance Inflation Factor (VIF) greater than five have been case by case removed (with some exceptions, some ratios with a VIF value no greater than ten have been kept for further testing since these are ratios that seem important in other research). In the end, 23 features have been removed. These VIF values have been chosen as these are general cutoff values (Craney & Surles, 2002).

In addition, the dataset is divided into 70% training data and 30% testing data. This has been done using the createDataPartition function of Caret to be able to test the model on unseen data. Finally, this study does not need to consider imbalance of the target variable, since the binary target variable is well distributed, as shown in Figure 3 below. At approximately 45% of the observations, the stock price goes down the following year, and at approximately 55% of the observations, the stock price goes up.

Figure 3: Distribution target variable Class.



3.4 *Feature selection*

After these steps, the total dataset consists of 19,809 observations and 119 features. Reducing features can prevent overfitting and reduce computation complexity (Raschka, 2022). In order to reduce the features, there are two commonly used methods, feature selection, and feature extraction. Feature extraction has not been applied since a disadvantage of feature extraction is that the new features are difficult to interpret, while sub-questions of this study also concern which indicators have an effect. The goals of feature selection are to improve data mining performance, to work with clean and understandable data, and to build simpler and more understandable models (J. Li et al., 2017). Therefore, feature selection is used, where the individual effects of the features are also interpretable.

Various feature selection methods are examined. Testing has been done with the Boruta package and filter methods. However, the wrapper method stepwise selection performs best. Therefore, this method has been applied. In the end, 75 features were removed by this wrapper method. Causing the final dataset for the models consists of 19,809 observations with 43 features and the target variable. All features of the final dataset can be found in Appendix C (page 35) Table 5.

3.5 *Algorithms*

This research will use seven machine learning algorithms to predict stock price movements, up or down, based on internal financial and macroeconomic indicators. These algorithms are tested to see which performs best. The literature section 2.5 extensively discusses why the different algorithms are used in this research. Below is an overview of which algorithms will be used:

- Decision Tree
- Gradient Boosting Machine
- k-Nearest Neighbors
- Logistic Regression
- Naive Bayes
- Random Forest
- Support Vector Machine

The algorithms are trained with the same dataset, the final dataset generated with the stepwise selection wrapper method. In addition, the features have been standardized. Firstly, standardizing features is important because there are features with different units. For example, some features with numbers such as Net income and EPS exist. However, the dataset also consists of ratios such as the P/E-ratio and Debt/Equity-ratio. In addition, it is also a requirement for specific machine learning algorithms to standardize the features (Raschka, 2014).

When training the algorithms, 10-fold cross-validation and random search from Caret are used to tune the hyperparameters. Despite cross-validation increasing the training time, it is implemented to see how the model works on new data, it reduces overfitting and allows the best hyperparameters to be selected (Berrar, 2019). Furthermore, 10-fold cross-validation has been used to tackle the problem of data leakage. In conclusion, this will lead to a robust model. Ultimately, the performances of the models are compared to the performances of the models on the test data. Random search has been implemented since the dataset consists of 43 features, and random search is efficient for hyperparameter tuning according to previous research (Bergstra & Bengio, 2012). Random search is chosen instead of Grid search as it reduces the training time slightly and to see other values in the distribution. Besides, Appendix D (page 36) shows the basic assumptions of the different algorithms used for this research.

Furthermore, the effects of the features are examined using algorithms that are well for interpretation purposes, for example, LR, but also using the varImp function of Caret. In addition, the correlations are examined. Finally, dummy features of the sectors have been added to see the difference between sectors.

3.6 *Evaluation Metrics*

Central to this thesis is a binary classification problem, whether the stock price goes up or down in the next year. A confusion matrix will be generated for the different models to see evaluation metrics. Firstly, accuracy will be used as an evaluation metric. Since the model can be used for investing, the false positives are more important than the false negatives. Therefore, the evaluation metrics precision, recall, F-score, and Area Under the Curve (AUC) will be used. Finally, the evaluation metric Kappa will be used. All these evaluation metrics can be obtained from the confusion matrix in the Caret package, except for the AUC. The package pROC was used for the AUC.

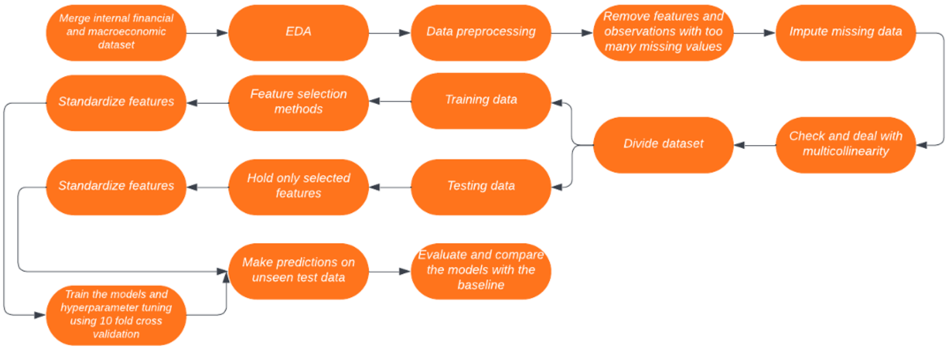
Although, as can be argued that the false positives are more important than the false negatives, the evaluation metric accuracy is used to select the optimal model. Accuracy has been chosen because there is a balanced target variable, and false negatives can also play an important role in trading strategies.

A baseline model will be used to assess the different models. The majority class classifier will be used as the baseline model. The final ground-truth labeled data does not need to be generated because this is already a variable in the dataset.

3.7 *Flowchart research*

The main steps within this research are shown in Figure 4 below. A few steps after the data preprocessing are also part of the preprocessing, but are mentioned separately, as they are of great importance within this research. What the various steps entail can be read in the previous sections, while the flowchart is visualized here. Besides, the CRISP-DM lifecycle approach will be used. Since there is much movement back and forth between the stages and it is not a linear process.

Figure 4: Flowchart research.


Merge internal financial and macroeconomic dataset
EDA
Data preprocessing
Remove features and observations with too many missing values
Impute missing data
Standardize features
Feature selection methods
Training data
Divide dataset
Check and deal with multicollinearity
Standardize features
Hold only selected features
Testing data
Make predictions on unseen test data
Evaluate and compare the models with the baseline
Train the models and hyperparameter tuning using 10 fold cross validation

4 RESULTS

This chapter discusses the results and will start with the baseline model. Furthermore, the hyperparameter tuning and how the different models perform in predicting stock price movements will be discussed. Finally, the effects of indicators will be discussed.

4.1 *Baseline model*

Firstly, the baseline model will be discussed since this is the model that will be used to assess the different models examined. The majority class classifier will be used as the baseline model. The majority class classifier looks at the largest class in the training data, and this class is always predicted as the outcome of the test data. This model has been used as a baseline model as it generates better accuracy than other random classifiers since it uses more information (Brownlee, 2019).

4.2 *Hyperparameter tuning*

In order to find the best hyperparameters, Random search is applied within the Caret package. Caret's modelLookup tool makes it possible to see which hyperparameters can be tuned (Elsinghors, 2022). Table 1 below shows which hyperparameters can be tuned in Caret for the different algorithms, what the hyperparameters to be tuned mean, and which values best performs and are ultimately used for the models.

Table 1: Overview hyperparameter tuning.

Algorithm	**Hyperparameter(s)**	**Optimal hyperparameter(s)**
Decision Tree	- Cp: minimum improvement in the model needed at each node, also known as complexity parameter (*Decision Trees in R*, 2022)	- Cp: 0.0022
Gradient Boosting Machine	- n.trees: Number of gradient boosting iterations (Greenwell, Boehmke, Cunningham, Developers, & Greenwell, 2019) - interaction.depth: Maximum depth of the trees (Greenwell et al., 2019) - shrinkage: Learning rate (Greenwell et al., 2019) - n.minobsinnode: Minimum number of observations in the trees' terminal nodes (Greenwell et al., 2019)	- n.trees: 4,492 - interaction.depth: 5 - shrinkage: 0.2286 - n.minobsinnode: 9
k-Nearest Neighbors	- K: number of nearest neighbors to include in the majority of the voting process (Subramanian, 2019)	- K: 27
Logistic Regression	- None	- None
Naive Bayes	- fL: To incorporate the Laplace smoother (*Naïve bayes classifier*, 2022) - usekernel: To use a kernel density estimate for continuous variables versus a gaussian density estimate (*Naïve bayes classifier*, 2022) - adjust: To adjust the bandwidth of the kernel density (*Naïve bayes classifier*, 2022)	- fL: 0 - usekernel: TRUE - adjust: 1
Random Forest	- Mtry: Number of features randomly sampled as candidates at each split (Brownlee, 2020).	- Mtry: 6
Support Vector Machine	- C: Penalty parameter (Yildirim, 2020) - Sigma: Controls the level of non-linearity (Theodoropoulos, 2020) - Kernel: take data as input and transform it (Theodoropoulos, 2020)	- C: 1 - Sigma: 0.05 - Kernel: Radial Basis Function

4.3 *Prediction of stock price movement*

This section discusses and compares the results of the different prediction models with the baseline model using the evaluation metrics. Table 2 below shows the evaluation metrics for the models on the test set. In addition, the baseline evaluation metrics can be seen to assess the results of the different models.

Table 2: Evaluation metrics of the different models on the test set.

Algorithm	Accuracy	Precision	Recall	F1-score	AUC	Kappa
Baseline	0.5415	0.5415	1.00	0.7025	0.50	0.00
Decision Tree	0.6882	0.7030	0.7345	0.7184	0.6841	0.3696
Gradient Boosting Machine	0.6694	0.6847	0.7219	0.7028	0.6647	0.3310
k-Nearest Neighbors	0.6664	0.6732	0.7462	0.7078	0.6592	0.3216
Logistic Regression	0.6850	0.6847	0.7751	0.7271	0.6768	0.3579
Naive Bayes	0.6297	0.6182	0.8266	0.7074	0.6119	0.2308
Random Forest	0.7024	0.7212	0.7343	0.7277	0.6995	0.3997
Support Vector Machine	0.6775	0.6878	0.7406	0.7132	0.6718	0.3460

As shown in Table 2, all models outperform the baseline model based on the evaluation metrics. The baseline has a higher evaluation metric only with the evaluation metric recall. This is the case since the majority class classifier predicts that all stock prices will increase (positive). Kappa takes the imbalance in the target variable into account, and therefore the kappa score of the baseline is zero. All kappa scores of the models are between 0.23 and 0.40 and can therefore be regarded as fair according to McHugh (2012). It can be concluded from the results that internal financial and macroeconomic indicators play a role in the classification problem of whether stock prices go up or down in the following year, as there are better evaluation scores than with the baseline model.

In addition, the RF model performs best. The RF model scores best on all evaluation metrics except for the evaluation metric recall. The NB model scores best on the evaluation metric recall, with a score of 82.66%. For most evaluation metrics, the NB model scores the worst of all models and comes closest to the baseline model. Based on the accuracy, it is striking that all models have at least an 8.82% higher accuracy than the baseline model. At the baseline model, the accuracy is 54.15%. The RF model even scores 16.09% higher, with a final accuracy score of 70.24%. Moreover, the evaluation scores are generally close to each other. For instance, looking at the accuracy scores and not including the baseline and NB model, the accuracy score of the best model (RF) and the worst model (KNN) only differ by 3.60%.

Furthermore, the results on the training set are very similar to the results on the test set. Table 3 below visualizes the accuracy on the train and test data and the difference for the models. Regarding accuracy, the biggest difference between the train and test data is 1.19%. To conclude, the differences are minimal, indicating that the models generalize well to new data and are not overfitting.

Table 3: Accuracy on the train and test set.

Algorithm	Train accuracy	Test accuracy	Difference
Decision Tree	69.37%	68.82%	0.54%
Gradient Boosting Machine	67.10%	66.94%	0.15%
k-Nearest Neighbors	67.84%	66.64%	1.19%
Logistic Regression	68.79%	68.50%	0.29%
Naive Bayes	62.60%	62.97%	-0.37%
Random Forest	71.13%	70.24%	0.89%
Support Vector Machine	68.27%	67.75%	0.52%

Figure 5 below shows the confusion matrix of the best-performing model, the RF model. The confusion matrix shows that 2,415 cases were correctly predicted when the stock price went down, and 3,150 cases were correctly predicted when the stock price went up. In total, 2,358 cases were incorrectly predicted, 29.76%. Furthermore, the figure shows that there is some difference in the quality of prediction between the two different classes. When the stock price goes up the following year, 26.57% is incorrectly predicted, and when the stock price goes down, 33.53% is incorrectly predicted. This distribution is wider for the other models, especially the NB model, which is therefore discussed below. It is striking that all models are better at predicting when the stock price goes up than when the stock price goes down. In addition, concerning the errors, there is a different error rate distribution in which year the stock price movement was predicted with the best-performing model. For example, the error rate in 2014 and 2016 was much higher than in the other years, 34.66% and 33.64%, respectively. While in the years 2015, 2017, and 2018 the error rate was 25.70%, 27.46%, and 27.68%, respectively. Besides, the various evaluation metrics of the confusion matrix can be seen in the figure.

Figure 5: Confusion Matrix Random Forest.

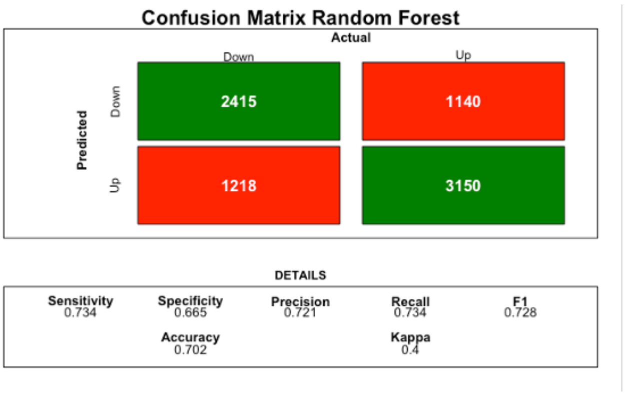
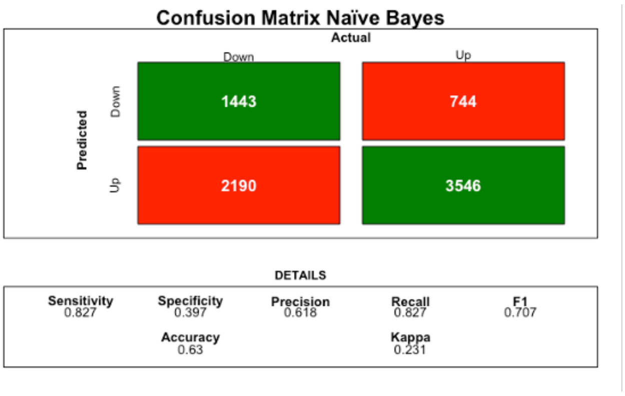


Figure 6 below shows the confusion matrix of the NB model. This model is further explained since the performance is worst on most evaluation metrics, except for the baseline model. However, the model performs best based on the evaluation metric recall. Recall is the ratio of correct positive predictions to the actual total positives. Compared to the other models, this model has a much larger spread in prediction accuracy between the two classes. When the stock price goes up, only 17.34% is misclassified. However, when the stock price goes down, 60.28% is misclassified. Consequently, this algorithm predicts many more cases of the stock price going up (72.40%) than going down (27.60%). The confusion matrices of the other models can be seen in Appendix E (page 37).

Figure 6: Confusion matrix Naïve Bayes.



To conclude, all models outperform the baseline model and have an accuracy between 63% and 70%. In addition, the RF model performs best and the NB model worst on most evaluation metrics. Furthermore, the results of the other algorithms are close to each other.

4.4 *Indicators influencing the stock price movement*

As discussed in the literature section, algorithms differ in their interpretability. Since this research also wants to examine the effect which internal financial and macroeconomic indicators influence the stock price movement, it is important to see the effects of the different features. This is not immediately visible with all algorithms. Therefore, the function varImp from Caret was used. Figure 7 below shows the fifteen most important features of the best-performing model, RF. In Appendix F (page 39), the same figures for the DT, LR, and GBM model are presented. There are no figures for the SVM, KNN, and NB algorithms since Caret's varImp function does not support these algorithms.

Figure 7: Feature importance Random Forest.

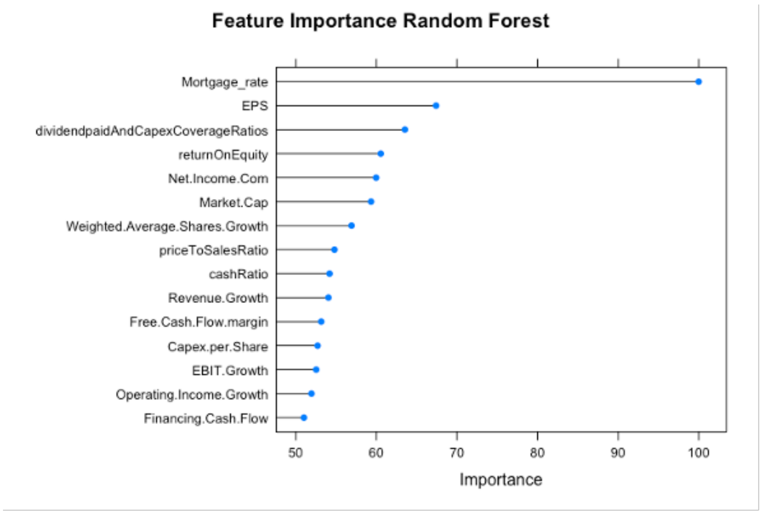


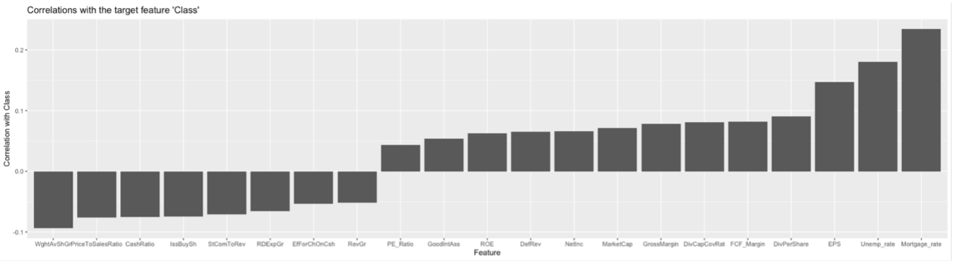
Figure 7 shows that the feature Mortgage rate is the most important feature of the RF model. Furthermore, the top five most important features consist of EPS, Dividend paid and capex coverage ratios, ROE, and Net income, respectively.

Considering the other algorithms in Appendix F (page 39), the Mortgage rate is the most important feature for all models. In addition, EPS is also number two of the most important features for the DT model and is in the top five features for the other models. Furthermore, the feature Dividend paid and capex coverage ratios also appear in the top six features in the DT and GBM model. In addition, the fourth most important feature of the RF model, the ROE, also appears in the top ten features of the GBM model. Finally, it is visible that Net income belongs to the five most important features in both the RF and DT model.

An algorithm that suits better for seeing relationships between features is LR. The output of the LR model can be seen in Appendix G (page 41), Table 6. The figure is divided into three feature types, internal financial indicators, macroeconomic indicators, and the dummy features of the various sectors. In addition, the features within these three types are sorted based on significance. It is striking in this figure that many of the features are significantly associated with the target variable. Firstly, the figure shows that all macroeconomic indicators have a highly significant effect. Secondly, several internal financial indicators have a highly significant effect, including, for example, the EPS, Gross margin, Deferred revenue, PE-ratio, ROE, Fixed asset turnover, and Debt/Equity-ratio. Finally, the effects for the sectors are significant, indicating that company's sector influences stock price movements.

Furthermore, the correlations between the various features and the target variable are examined. Figure 8 below shows the features that have a correlation higher than 0.05 or lower than -0.05, the top 20 features in terms of correlation. The highest correlation is 0.23, which is generally considered a weak correlation. To summarize, many features seem to correlate with the target variable, but this correlation is low. However, many of the same features emerge here as in the feature importance plots of the various models.

Figure 8: Highest correlated features.



To conclude, it is visible from the feature importance plots, results of the LR model, and the correlations that macroeconomic indicators, such as the Mortgage rate, play a major and higher role in stock price movements the following year than the various internal financial indicators. However, it is visible that various internal financial indicators affect stock price movements, but this effect is not very high. Besides, it is striking that the ratios and growth measures are more present as important features than different figures such as revenue. This can be explained by the fact that ratios consider the different sizes of companies. For example, a revenue of 100,000 euros for a large company does not mean much, but it does mean a lot for a small company.

5 DISCUSSION

This chapter starts with a summary and discussion of the results. Subsequently, the results will be compared with the literature. Thirdly, the scientific and societal impact of this research will be discussed. Lastly, the limitations and future directions of this research will be discussed.

5.1 *Summary and discussion of the results*

The purpose of this research was to predict whether stock prices will increase or decrease the following year based on internal financial and macroeconomic indicators that are publicly available. Therefore, the following research question was central to this study:

> *To what extent is it possible to predict stock price movements in the next year based on last year's internal financial and macroeconomic indicators?*

To fulfill this research question, several sub-questions are needed. The following sub-questions were addressed in this research to answer the central question:

RQ1 *Which internal financial indicators are important in the binary prediction of stock price movements?*

RQ2 *Which macroeconomic indicators are important in the binary prediction of stock price movements?*

RQ3 *Which machine learning algorithm can best predict binary stock price movements?*

All models outperform the baseline model based on the different evaluation metrics. Therefore, it can be concluded that internal financial and macroeconomic indicators play a role in the classification problem of whether the stock price goes up or down in the following year. Based on the accuracy, all models have at least an 8.82% higher accuracy than the baseline model. The RF model performs best based on the different evaluation metrics. The RF model even scores 16% higher than the baseline model based on accuracy, with a final accuracy score of 70.24%. For most evaluation metrics, the NB model scores the worst of all models and comes closest to the baseline model in evaluation scores. Moreover, the evaluation scores of the different models are close, and all models are better at predicting when the stock price goes up than when the stock price goes down.

When looking at the individual effects of internal financial and macroeconomic indicators on stock price movements, it is striking that macroeconomic indicators, such as the Mortgage rate, play a major and higher role in stock price movements than the various internal financial indicators. It is also visible that various internal financial indicators affect stock price movements, such as the EPS, Dividend paid and capex coverage ratios, ROE, Net income, PE-ratio, Fixed asset turnover, and Debt/Equity-ratio.

5.2 *Comparison to the literature*

The literature has many different results about which algorithm best predicts binary stock price movements. The result that the RF algorithm performs best is in line with Basak et al. (2019); Illa et al. (2022). However, the results are contradictory to Ravikumar and Saraf (2020) research, where the SVM, LR, and NB algorithms scored better in accuracy than the RF algorithm. In addition, the results of the different algorithms are close, which is also reflected in the literature. Various studies have different algorithms as the best performing algorithm. According to Hu et al. (2013), the SVM algorithm is an excellent algorithm for predicting stock prices, while according to Subha and Nambi (2012), the KNN algorithm performs very well. In line with Hutapea et al. (2019), based on accuracy, the DT model performs better than the NB model within this study. Furthermore, it is interesting that Kumar et al. (2018) used four of the same algorithms as this study, KNN, NB, RF, and SVM. In line with this research, the RF model has been found to perform best with large datasets.

When investigating the effect of internal financial indicators on stock price movements, this study has results that are in line with, but also contradictory to, previous studies. The conclusions in previous studies vary widely as to which internal financial indicators play a role. For example, this research is in line with Ali et al. (2018); Banchuenvijit (2016), that the Debt/Equity-ratio plays a role in stock price movements prediction. Furthermore, this study aligns with Ali et al. (2018); Arkan (2016); Puspitaningtyas (2017) that the EPS plays a role in stock price movement prediction. In addition, this research is in line with Öztürk and Karabulut (2018) that the P/E-ratio significantly positively affects stock prices. Contrary to Banchuenvijit (2016); Jiang et al. (2019); Vedd and Yassinski (2015), this research found no significant effect on the Total asset turnover.

However, there are also contradicting results in previous studies. According to Ali et al. (2018); Banchuenvijit (2016), the CR has a significant effect on stock price movements, while this is not the case in this study, which is in line with Öztürk and Karabulut (2018); Puspitaningtyas (2017); Vedd and Yassinski (2015); Yuliarti and Diyani (2018). The same applies to the NPM, Banchuenvijit (2016); Öztürk and Karabulut (2018) found significant effects, while this study finds no significant effects, which is in line with Vedd and Yassinski (2015). Besides, this research is in line with Ali et al. (2018); Arkan (2016), regarding that the ROE significantly affects stock price movements, which is contradictory to the results of Puspitaningtyas (2017); Yuliarti and Diyani (2018), where the ROE has no significant effect. In line with Kwag and Kim (2013), this research finds industry effects.

Regarding the effects of internal financial indicators, there are more internal financial indicators in this study that influence stock price movements, such as Dividend paid and capex coverage ratios and Net income. These internal financial indicators have often not been covered in previous studies, as this study consisted of a very large dataset of internal financial indicators, while earlier studies often consisted of a much smaller dataset.

The macroeconomic indicators play a major role in predicting stock price movements within this study. This is in line with Abbas et al. (2019); Chen (2009); Ma et al. (2022). Of all features, the Mortgage rate is the most important. This aligns with Ma et al. (2022), where the housing market is an important component in predicting stock price movements.

5.3 *Discussion of scientific and societal impact*

There is a wide range of literature on which internal financial and macroeconomic indicators influence stock price movements. However, this study is different because a prediction model is built, which has not been done much in previous studies. Secondly, this research differs from previous studies since it does not predict the exact daily stock price movement. However, a binary classification problem is applied whether the stock price goes up or down the following year. Thirdly, seven algorithms were tested in this study, with the results being reasonably close. Fourthly, this study uses both internal financial and macroeconomic indicators. Many previous studies used one of the two indicators, but the effect together has been less studied. Lastly, an extensive dataset (more than 200 features) was used for this study, which eventually resulted in 43 features used for the final models. In previous studies, the feature set was often much more limited.

This research aligns with previous research on which specific internal financial and macroeconomic indicators have an effect. However, due to the extensive feature set, new internal financial indicators have been discovered that play a significant role in stock price movements. In conclusion, from previous studies, the CR, Debt/Equity-ratio, EPS, ROE, and Asset turnover ratio were especially important. This study confirmed these results, except for the CR. However, for example, the Dividend paid and capex coverage ratios, Net income, Market cap, Weighted average shares growth, Price to sales ratio, Cash ratio, and Revenue growth are also in the top 10 features in terms of importance for the best-performing model (RF).

Regarding societal relevance, this research is about building a model with the best-performing features and algorithm to distinguish stocks that go up in price from stocks that go down in price. This was the aim with the ultimate purpose that individuals with less insight into the market can also set up a trading strategy, as the data used is publicly available. It certainly succeeded in creating a model that helps people to invest, as the model predicts 16% better than the majority class classifier. However, the accuracy is 70%, meaning the model cannot predict all stock price movements correctly.

5.4 *Limitations and future directions*

Besides the strengths of the study, there are some limitations. Firstly, the dataset has many missing values. When a feature has more than 15% missing cases, this feature has been dropped, and when an observation has more than 10% missing features, this observation is dropped. Consequently, 60 features have been removed from the dataset and are therefore not examined. Secondly, no data from the past few years has been examined. The last year for which stock price movements were examined was 2019, as this is the last year the dataset has been published. It will therefore be relevant for future research to include more recent years. Especially nowadays with the war between Ukraine and Russia and the COVID crisis in recent years. This might impact stock price movements and, therefore, ultimately, the trading strategy for individuals. Thirdly, in terms of the practical relevance of this study, there is another limitation. Since this research is a binary classification problem, it only distinguishes the stocks that go up in price from those that go down in price. However, it is also interesting for an investor to see how much the stock price goes down or up, allowing an investor, for example, to distinguish 'excellent stocks' from 'good stocks'. This could be the subject of future research and would optimize the trading strategy even more. Fourthly, a limitation of the scope of this study is that it is limited only to the effects of various internal financial and macroeconomic indicators on stock price movements. However, no further explanation is given as to why certain features have an effect. Therefore, causality is not considered. Lastly, this research only uses machine learning algorithms and no deep learning methods. However, for instance, Zhong and Enke (2017) concluded that ANN works better than LR in predicting stock price movements. That is why it might be interesting to test deep learning methods in predicting stock price movements in the future.

To conclude, research in the future can be done with a more recent and complete dataset. In addition, deep learning methods and a regression problem can be used and there can be more focus on causality.

6 CONCLUSION

This research aimed to classify stocks that go up in price and those that go down in price to set up a trading strategy for individuals with less knowledge of the market. This has been done utilizing a binary classification problem, using internal financial and macroeconomic indicators as predictors. This goal has led to the following research question:

> *To what extent is it possible to predict stock price movements in the next year based on last year's internal financial and macroeconomic indicators?*

To fulfill this research question, three topics are essential. The effect of internal financial indicators on stock price movements, the effect of macroeconomic indicators, and which machine learning algorithms best predict binary stock price movements. Therefore, these aspects have been discussed in the literature. Concerning the macroeconomic indicators, all studies agree that they play an essential role in stock price movements. For example, Inflation, Unemployment, and the Housing market are important indicators. Previous studies have shown different results regarding which internal financial indicators play a role. However, the CR, Debt/Equity-ratio, EPS, ROE, and Asset turnover ratio have a significant effect in multiple studies.

After the literature section, the research was put into practice. Seven machine learning algorithms that appeared to be important in the literature, DT, RF, GBM, SVM, KNN, NB, and LR, were used to answer the research question. The different algorithms were trained on a dataset of 43 features, 19,809 observations of US companies, and the target variable, whether the stock price went up or down the following year. The results are in line with the literature. Macroeconomic indicators such as the Mortgage rate play an important role. In addition, the results of this research align with previous research on which internal financial have an effect, except for the CR. However, due to the extensive feature set, new internal financial indicators have also been discovered that play a role in stock price movements. For example, the features Dividend paid and capex coverage ratios, Net income, Market cap, Weighted average shares growth, Price to sales ratio, Cash ratio, and Revenue growth are in the top 10 features in terms of importance for the best-performing model, RF.

All models used in this study outperform the baseline model. In addition, the performances of the different models are close. The best-performing model, RF, has an accuracy of 70.24%, which is 16.09% higher than the baseline model. In conclusion, internal financial and macroeconomic indicators play a reasonably large role in binary stock price movements the following year. Because of this research, a model now allows investors with less knowledge of the market to distinguish better stocks that go up in price from stocks that go down in price.

Future research can examine recent years, and deep learning methods can be implemented to see if the results differ. In addition, a regression problem can be used instead of a binary classification problem to distinguish the quality of stocks in which investments are made even more. Finally, future research may focus on the causality between internal financial and macroeconomic indicators and stock price movements.

REFERENCES

Abbas, G., Hammoudeh, S., Shahzad, S. J. H., Wang, S., & Wei, Y. (2019). Return and volatility connectedness between stock markets and macroeconomic factors in the g-7 countries. *Journal of Systems Science and Systems Engineering*, *28*(1), 1–36.

Ali, S. S., Mubeen, M., Lal, I., & Hussain, A. (2018). Prediction of stock performance by using logistic regression model: evidence from pakistan stock exchange (psx). *Asian Journal of Empirical Research*, *8*(7), 247–258.

Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M. K. A. (2013). Stock price prediction using k-nearest neighbor (knn) algorithm. *International Journal of Business, Humanities and Technology*, *3*(3), 32–44.

Andri et mult. al., S. (2022). DescTools: Tools for descriptive statistics [Computer software manual]. Retrieved from https://cran.r-project.org/package=DescTools (R package version 0.99.47)

Arkan, T. (2016). The importance of financial ratios in predicting stock price trends: A case study in emerging markets. *Finanse, Rynki Finansowe, Ubezpieczenia*, *79*, 13–26.

Aspara, J., & Tikkanen, H. (2011). Individuals' affect-based motivations to invest in stocks: Beyond expected financial returns and risks. *Journal of Behavioral Finance*, *12*(2), 78–89.

Banchuenvijit, W. (2016). Financial ratios and stock prices: Evidence from the agriculture firms listed on the stock exchange of thailand. *UTCC International Journal of Business & Economics*, *8*(2), 23–29.

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

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Berrar, D. (2019). *Cross-validation.*

Brownlee, J. (2019, Sep). *How to develop and evaluate naive classifier strategies using probability.* Retrieved from https://machinelearningmastery.com/how-to-develop-and-evaluate-naive-classifier-strategies-using-probability/

Brownlee, J. (2020, Jul). *Tune machine learning algorithms in r (random forest case study).* Retrieved from https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/

Carbone, N. (2019). *200+ financial indicators of us stocks (2014-*

2018). Retrieved from https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018

Chen, S.-S. (2009). Predicting the bear stock market: Macroeconomic variables as leading indicators. *Journal of Banking & Finance*, *33*(2), 211–223.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality engineering*, *14*(3), 391–403.

Dash, R., Samal, S., Rautray, R., & Dash, R. (2019). A topsis approach of ranking classifiers for stock index price movement prediction. In *Soft computing in data analytics* (pp. 665–674). Springer.

Decision trees in r. (2022). Retrieved from https://www.learnbymarketing.com/tutorials/rpart-decision-trees-in-r/#:~:text=cp%3A%20Complexity%20Parameter,misclassification%20at%20every%20terminal%20node.

Elsinghors, S. (2022). *Hyperparameter tuning in caret: R.* Retrieved from https://campus.datacamp.com/courses/hyperparameter-tuning-in-r/introduction-to-hyperparameters?ex=8

Fisher, A., Martineau, C., & Sheng, J. (2022). Macroeconomic attention and announcement risk premia. *The Review of Financial Studies*, *35*(11), 5057–5093.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., … Graves, S. (2012). Package 'car'. *Vienna: R Foundation for Statistical Computing*, *16*.

Freedman, V. A., & Wolf, D. A. (1995). A case study on the use of multiple imputation. *Demography*, *32*(3), 459–470.

Fromentin, V., Lorraine, M., Ariane, C., & Alshammari, T. (2022). Time-varying causality between stock prices and macroeconomic fundamentals: Connection or disconnection? *Finance Research Letters*, *49*, 103073.

Greenwell, B., Boehmke, B., Cunningham, J., Developers, G., & Greenwell, M. B. (2019). Package 'gbm'. *R package version*, *2*(5).

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *6*, e5518.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of finance and data science*, *4*(3), 183–201.

Heo, J., & Yang, J. Y. (2016). Stock price prediction based on financial statements using svm. *International Journal of Hybrid Information Technology*, *9*(2), 57–66.

Hu, Z., Zhu, J., & Tse, K. (2013). Stocks market prediction using support vector machine. In *2013 6th international conference on information* *management, innovation management and industrial engineering* (Vol. 2, pp. 115–118).

Hutapea, J. Y., Samuel, Y. T., & Sitorus, H. (2019). Comparison of accuracy between two methods: Nave bayes algorithm and decision tree-j48 to predict the stock price of pt astra international tbk using data from indonesia stock exchange. In *Abstract proceedings international scholars conference* (Vol. 7, pp. 1244–1258).

Illa, P. K., Parvathala, B., & Sharma, A. K. (2022). Stock price prediction methodology using random forest algorithm and support vector machine. *Materials Today: Proceedings*, *56*, 1776–1782.

Jiang, Q., Wang, X., Li, Y., Wang, D., & Huang, Q. (2019). Financial indicators and stock price movements: The evidence from the finance of china. In *International conference on management science and engineering management* (pp. 743–758).

Kaplan, J. (2020). fastdummies: Fast creation of dummy (binary) columns and rows from categorical variables [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/fastDummies/index.html (R package version 1.6.3)

Karakus, R., & Bozkurt, I. (2017). The effect of financial ratios and macroeconomic factors on firm value: An empirical analysis in borsa istambul. In *Rsep international conferences on social issues and economic studies* (Vol. 4).

Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & Mohammad, J. R. (2014). Advantage and drawback of support vector machine functionality. In *2014 international conference on computer, communications, and control technology (i4ct)* (pp. 63–65).

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, *28*, 1–26.

Kuhn, M., Jackson, S., & Cimentada, J. (2022). corrr: Correlations in r [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/corrr/index.html (R package version 0.4.4)

Kumar, I., Dogra, K., Utreja, C., & Yadav, P. (2018). A comparative study of supervised machine learning algorithms for stock market trend prediction. In *2018 second international conference on inventive communication and computational technologies (icicct)* (pp. 1003–1007).

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of statistical software*, *36*, 1–13.

Kwag, S. W., & Kim, Y. S. (2013). Stock price predictability of financial ratios and macroeconomic variables: A regulatory perspective. *Industrial Engineering and Management Systems*, *12*(4), 406–415.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys*

(CSUR), *50*(6), 1–45.

Li, Z., & Tam, V. (2017). A comparative study of a recurrent neural network and support vector machine for predicting price movements of stocks of different volatilites. In *2017 ieee symposium series on computational intelligence (ssci)* (pp. 1–8).

Ma, F., Lu, X., Liu, J., & Huang, D. (2022). Macroeconomic attention and stock market return predictability. *Journal of International Financial Markets, Institutions and Money*, *79*, 101603.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Mendekar, V. (2021). *Machine learning - it's all about assumptions.* Retrieved from https://www.kdnuggets.com/2021/02/machine-learning-assumptions.html

Mirashi, S. (2022, Jul). *Us macroeconomic data.* Retrieved from https://www.kaggle.com/datasets/sarthmirashi07/us-macroeconomic-data

Naïve bayes classifier. (2022). UC Business Analytics R Programming Guide. Retrieved from https://uc-r.github.io/naive_bayes

Öztürk, H., & Karabulut, T. A. (2018). The relationship between earnings-to-price, current ratio, profit margin and return: an empirical analysis on istanbul stock exchange. *Accounting and Finance Research*, *7*(1), 109–115.

Pal, M., & Mather, P. M. (2001). Decision tree based classification of remotely sensed data. In *22nd asian conference on remote sensing* (Vol. 5, p. 9).

Puspitaningtyas, Z. (2017). Is financial performance reflected in stock prices? In *2nd international conference on accounting, management, and economics 2017 (icame 2017)* (pp. 17–28).

Raschka, S. (2014). About feature scaling and normalization and the effect of standardization for machine learning algorithms. *Polar Political Legal Anthropology Rev*, *30*(1), 67–89.

Raschka, S. (2022). *How do you attack a machine learning problem with a large number of features?* Retrieved from https://sebastianraschka.com/faq/docs/large-num-features.html

Ravikumar, S., & Saraf, P. (2020). Prediction of stock prices using machine learning (regression, classification) algorithms. In *2020 international conference for emerging technology (incet)* (pp. 1–5).

Reddy, V. K. S. (2018). Stock market prediction using machine learning. *International Research Journal of Engineering and Technology (IRJET)*, *5*(10), 1033–1035.

Revelle, W., & Revelle, M. W. (2015). Package 'psych'. *The comprehensive R archive network*, *337*, 338.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package 'mass'. *Cran r*, *538*, 113–120.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, *12*(1), 1–8.

Ryan, J. A., Ulrich, J. M., Thielen, W., Teetor, P., Bronder, S., & Ulrich, M. J. M. (2015). Package 'quantmod'. *Cran r*.

Schafer, J. L. (1998). The practice of multiple imputation. In *meeting of the methodology center, pennsylvania state university, university park, pa.*

Spinu, M. V. (2016). Package 'lubridate'. *Recuperado el*.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099–1104.

Subha, M., & Nambi, S. T. (2012). Classification of stock index movement using k-nearest neighbours (k-nn) algorithm. *WSEAS Transactions on Information Science & Applications*, *9*(9), 261–270.

Subramanian, D. (2019, Jun). *A simple introduction to k-nearest neighbors algorithm.* Towards Data Science. Retrieved from https://towardsdatascience.com/a-simple-introduction-to-k-nearest-neighbors-algorithm-b3519ed98e#:~:text='k'%20in%20KNN%20is%20a,majority%20of%20the%20voting%20process.

Theodoropoulos, C. (2020, Dec). *Support vector machines under the hood.* Towards Data Science. Retrieved from https://towardsdatascience.com/support-vector-machines-under-the-hood-c609e57a4b09

Tierney, N., Cook, D., McBain, M., & Fay, C. (2021). naniar: Data structures, summaries, and visualisations for missing data [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/naniar/index.html (R package version 0.6.1)

Times, T. E. (2022). *What is stock market?* Retrieved from https://economictimes.indiatimes.com/definition/stock-market

Tuszynski, J., & Khachatryan, M. H. (2013). *Package 'catools'.* Recuperado.

Ulandari, N. W. J., & Damayanthi, I. G. A. E. (2021). Macro-economic factors and financial ratios on stocks returns. *International Journal of Management and Commerce Innovations*, *9*, 1–10.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, *45*, 1–67.

Vedd, R., & Yassinski, N. (2015). The effect of financial ratios, firm size & operating cash flows on stock price: Evidence from the latin america industrial sector. *Journal of Business and Accounting*, *8*(1), 15.

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naive bayes. *Encyclopedia of machine learning*, *15*, 713–714.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). Package 'corrplot'. *Statistician*, *56*(316), e24.

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version*, *2*(1), 1–189.

Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from https://cran.r-project.org/web/packages/dplyr/index.html (R package version 1.0.10)

Wilcox, R. (2005). Trimming and winsorization. *Encyclopedia of biostatistics*, *8*.

Wilke, C. O., Wickham, H., & Wilke, M. C. O. (2019). Package 'cowplot'. *Streamlined Plot Theme and Plot Annotations for 'ggplot2*.

Yildirim, S. (2020, May). *Hyperparameter tuning for support vector machines - c and gamma parameters.* Towards Data Science. Retrieved from https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167

Yuliarti, A., & Diyani, L. A. (2018). The effect of firm size, financial ratios and cash flow on stock return. *The Indonesian Accounting Review*, *8*(2), 226–240.

Zhang, X. (1999). Using class-center vectors to build support vector machines. In *Neural networks for signal processing ix: Proceedings of the 1999 ieee signal processing society workshop (cat. no. 98th8468)* (pp. 3–11).

Zhong, X., & Enke, D. (2017). A comprehensive cluster and classification mining procedure for daily stock market return forecasting. *Neurocomputing*, *267*, 152–168.

7 APPENDIX A: DATA SOURCE/CODE/ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data. The code used in this thesis is not publicly available. There are no images within this research that the author did not produce.

8 APPENDIX B: USED PACKAGES

Table 4 below shows which packages were used for this research and for what purpose.

Table 4: Overview used packages.

Package	Purpose
Boruta (Kursa & Rudnicki, 2010)	Feature selection
Car (Fox et al., 2012)	To deal with multicollinearity
Caret (Kuhn, 2008)	Training, tuning and testing the different models
CaTools (Tuszynski & Khachatryan, 2013)	To deal with multicollinearity
Corrplot (Wei et al., 2017)	Plot correlation plot
Corrr (Kuhn, Jackson, & Cimentada, 2022)	For creating a data frame of correlations with the target variable
Cowplot (Wilke, Wickham, & Wilke, 2019)	Visualizing multiple figures in one single figure
DescTools (Andri et mult. al., 2022)	Winsorizing features
Dplyr (Wickham, François, Henry, & Müller, 2022)	Manipulating dataframes
fastDummies (Kaplan, 2020)	Convert a categorical feature into dummy variables
Ggplot2 (Wickham, Chang, & Wickham, 2016)	Explanatory data analysis and plotting results
Lubridate (Spinu, 2016)	Dealing with dates
Mass (Ripley et al., 2013)	Feature selection and to deal with multicollinearity
Mice (Van Buuren & Groothuis-Oudshoorn, 2011)	Multiple imputation for the missing values
Naniar (Tierney, Cook, McBain, & Fay, 2021)	Visualizing missing values
pROC (Robin et al., 2011)	Display the ROC curve and calculate AUC
Psych (Revelle & Revelle, 2015)	To calculate Cohen's Kappa
Quantmod (Ryan et al., 2015)	To deal with multicollinearity
Scales (Wickham et al., 2016)	To make distribution plot of target variable

9 APPENDIX C: OVERVIEW FEATURES FINAL DATASET

Table 5 below shows all features of the final dataset that are used for the models.

Table 5: Overview features final dataset.

Feature	**Type**
Class	Target variable
Revenue.Growth	Internal financial indicator
Net.Income...Non.Controlling.int	Internal financial indicator
Net.Income.Com	Internal financial indicator
EPS	Internal financial indicator
Dividend.per.Share	Internal financial indicator
Gross.Margin	Internal financial indicator
Free.Cash.Flow.margin	Internal financial indicator
Inventories	Internal financial indicator
Goodwill.and.Intangible.Assets	Internal financial indicator
Deferred.revenue	Internal financial indicator
Issuance..buybacks..of.shares	Internal financial indicator
Financing.Cash.Flow	Internal financial indicator
Effect.of.forex.changes.on.cash	Internal financial indicator
priceToSalesRatio	Internal financial indicator
priceEarningsRatio	Internal financial indicator
returnOnEquity	Internal financial indicator
fixedAssetTurnover	Internal financial indicator
cashRatio	Internal financial indicator
debtEquityRatio	Internal financial indicator
dividendpaidAndCapexCoverageRatios	Internal financial indicator
Market.Cap	Internal financial indicator
Intangibles.to.Total.Assets	Internal financial indicator
Stock.based.compensation.to.Revenue	Internal financial indicator
Days.Sales.Outstanding	Internal financial indicator
Capex.per.Share	Internal financial indicator
EBIT.Growth	Internal financial indicator
Operating.Income.Growth	Internal financial indicator
Weighted.Average.Shares.Growth	Internal financial indicator
R.D.Expense.Growth	Internal financial indicator
CPI	Macroeconomic indicator
Mortgage_rate	Macroeconomic indicator
Unemp_rate	Macroeconomic indicator
Year	Macroeconomic indicator
Sector_Basic_Materials	Sector indicator
Sector_Communication_Services	Sector indicator
Sector_Consumer_Cyclical	Sector indicator
Sector_Consumer_Defensive	Sector indicator
Sector_Energy	Sector indicator
Sector_Financial_Services	Sector indicator
Sector_Healthcare	Sector indicator
Sector_Industrials	Sector indicator
Sector_Real_Estate	Sector indicator
Sector_Technology	Sector indicator

10 APPENDIX D: ASSUMPTIONS USED ALGORITHMS

Below the basic assumptions of the different algorithms used for this research are stated.

- Decision Tree
  - At the start, whole training data is considered as root (Pal & Mather, 2001)
  - Records distributed recursively based on the attribute value (Pal & Mather, 2001)
- k-Nearest Neighbors
  - Data is in feature space (Mendekar, 2021)
  - Desirable to have k as an odd number in a binary classification problem (Mendekar, 2021)
- Logistic Regression
  - Minimal or no multicollinearity (Stoltzfus, 2011)
  - Needs a large sample size (Stoltzfus, 2011)
  - Observations are independent of each other (Stoltzfus, 2011)
- Naive Bayes
  - Conditional independence (Webb, Keogh, & Miikkulainen, 2010)
- Random Forest
  - No formal distributions (Hengl, Nussbaum, Wright, Heuvelink, & Gräler, 2018)
- Support Vector Machine
  - Data is independent and identically distributed (Zhang, 1999)

11 APPENDIX E: CONFUSION MATRICES MODELS

The confusion matrices of the RF and NB model are already shown in Section 4.3. Figures 9, 10, 11, 12, and 13 below show the confusion matrices, including the different evaluation metrics of the other algorithms, namely DT, GBM, KNN, LR, and SVM.

Figure 9: Confusion Matrix Decision Tree.

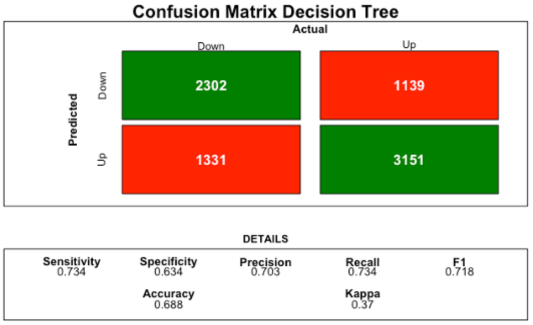


Figure 10: Confusion Matrix Gradient Boosting Machine.

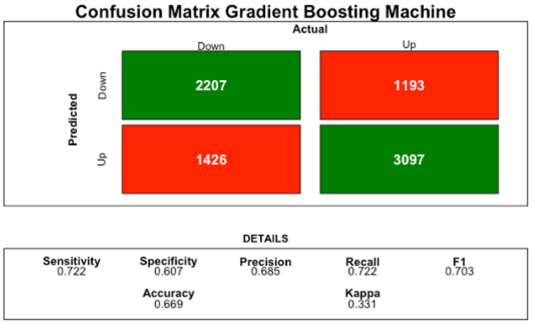


Figure 11: Confusion Matrix k-Nearest Neighbors.

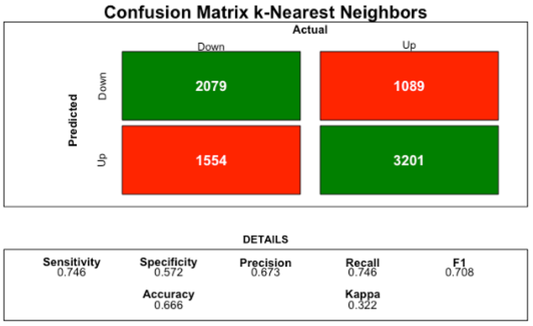


Figure 12: Confusion Matrix Logistic Regression.

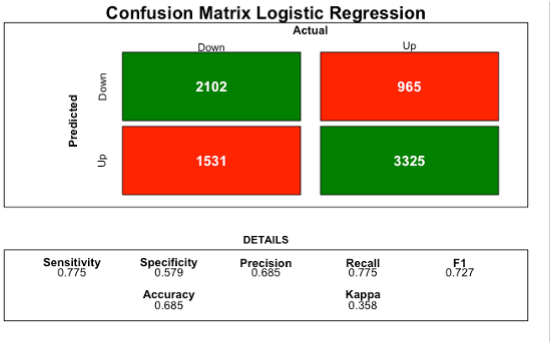
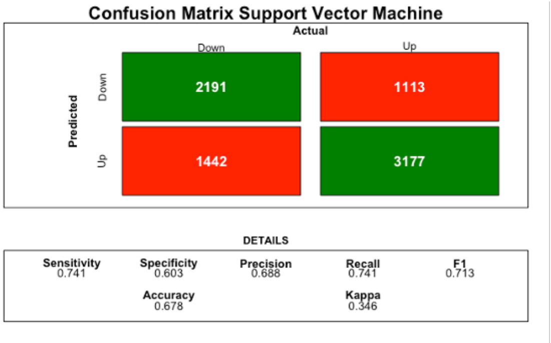


Figure 13: Confusion Matrix Support Vector Machine.



12 APPENDIX F: FEATURE IMPORTANCE MODELS

The feature importance plot of the RF model is already shown in Section 4.4. Figures 14, 15, and 16 below show the feature importance plots of the other algorithms where it is possible to create a feature importance plot, namely DT, GBM, and LR.

Figure 14: Top 15 features Decision Tree.

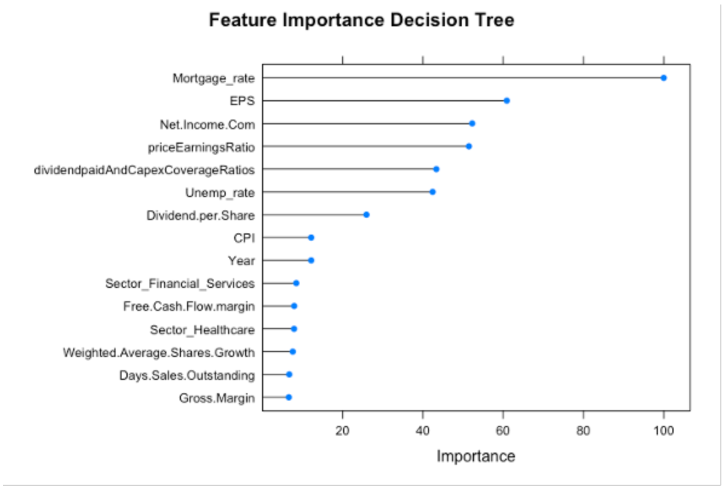


Figure 15: Top 15 features Gradient Boosting Machine.

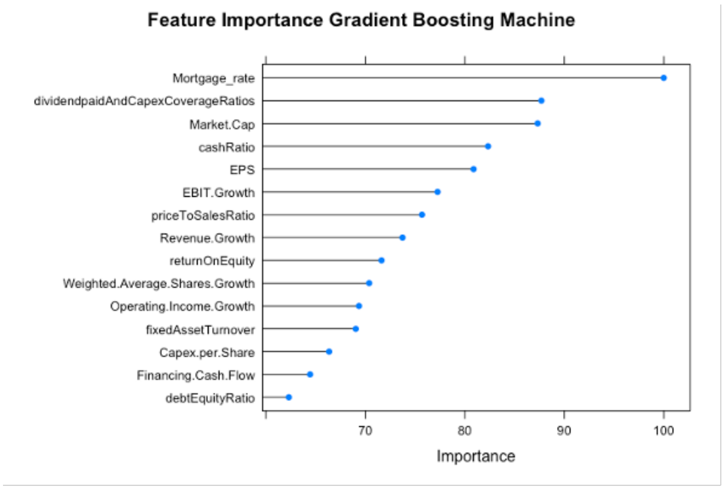
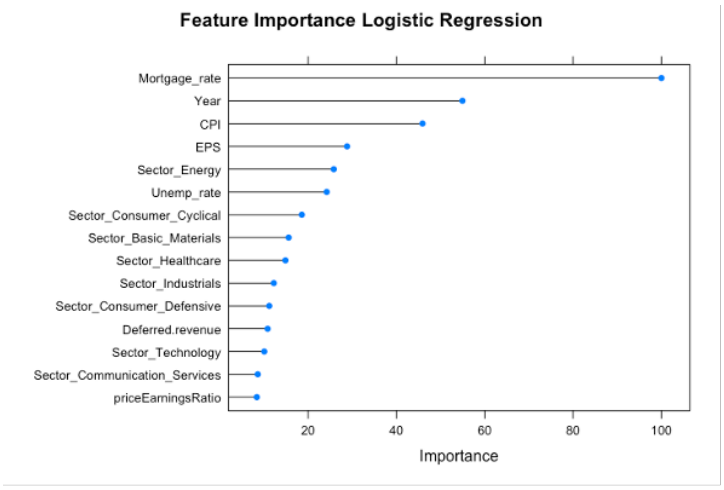


Figure 16: Top 15 features Logistic Regression.



13 APPENDIX G: RESULTS LOGISTIC REGRESSION MODEL

Table 6 below shows the results of the LR model.

Table 6: Output logistic regression model.

0	‘***’
0.001	‘**’
0.01	‘*’
0.05	‘.’
1	‘ ’

Feature type	Feature	Estimate	Std. error	Z value	Pr(>|z|)	Signif
	(Intercept)	0.21040	0.02038	10.325	<2e-16	***
Internal financial indicators	EPS	0.26291	0.02864	9.180	<2e-16	***
	Gross.Margin	0.08776	0.02609	3.363	0.000771	***
	Deferred.revenue	0.10221	0.02420	4.223	2.41e-05	***
	priceEarningsRatio	0.07582	0.02137	3.548	0.000389	***
	Inventories	0.06956	0.02685	2.591	0.009571	**
	Effect.of.forex.changes.on.cash	-0.06106	0.02263	-2.698	0.006973	**
	returnOnEquity	0.05827	0.02156	2.702	0.006890	**
	fixedAssetTurnover	-0.05672	0.02143	-2.647	0.008113	**
	debtEquityRatio	-0.06832	0.02086	-3.276	0.001053	**
	Market.Cap	0.13603	0.04522	3.008	0.002628	**
	Intangibles.to.Total.Assets	0.07816	0.02448	3.193	0.001409	**
	Days.Sales.Outstanding	-0.06818	0.02153	-3.167	0.001539	**
	Operating.Income.Growth	-0.08218	0.02939	-2.796	0.005174	**
	Weighted.Average.Shares.Growth	-0.06882	0.02316	-2.971	0.002969	**
	Revenue.Growth	-0.05418	0.02188	-2.476	0.013279	*
	Net.Income.Com	-0.08392	0.03998	-2.099	0.035807	*
	Dividend.per.Share	0.05832	0.02386	2.444	0.014516	*
	Goodwill.and.Intangible.Assets	-0.07133	0.03385	-2.107	0.035076	*
	Financing.Cash.Flow	0.05492	0.02495	2.201	0.027747	*
	priceToSalesRatio	-0.07549	0.03643	-2.072	0.038246	*
	Capex.per.Share	0.04279	0.02152	1.989	0.046743	*
	EBIT.Growth	0.06000	0.02950	2.034	0.041956	*
	Net.Income...Non.Controlling.int	-0.04136	0.02297	-1.800	0.071836	.
	Free.Cash.Flow.margin	0.07466	0.04932	1.514	0.130058	
	Issuance..buybacks..of.shares	-0.04339	0.03040	-1.428	0.153422	
	cashRatio	-0.03490	0.02317	-1.506	0.131957	
	dividendpaidAndCapexCoverageRatios	0.03116	0.02306	1.351	0.176648	
	Stock.based.compensation.to.Revenue	0.05955	0.04811	1.238	0.215727	
	R.D.Expense.Growth	-0.03382	0.02254	-1.500	0.133549	
Macro-economic indicators	CPI	-0.42593	0.03066	-13.891	<2e-16	***
	Mortgage_rate	101.676	0.03529	28.812	<2e-16	***
	Unemp_rate	0.21875	0.02765	7.912	2.53e-15	***
	Year	-0.50272	0.03067	-16.389	<2e-16	***
Sectors	Sector_Basic_Materials	-0.21626	0.03907	-5.535	3.11e-08	***
	Sector_Communication_Services	-0.09989	0.02768	-3.608	0.000308	***
	Sector_Consumer_Cyclical	-0.30876	0.04856	-6.358	2.05e-10	***
	Sector_Consumer_Defensive	-0.15050	0.03481	-4.324	1.53e-05	***
	Sector_Energy	-0.31119	0.03725	-8.355	<2e-16	***
	Sector_Healthcare	-0.30039	0.05631	-5.335	9.56e-08	***
	Sector_Industrials	-0.23338	0.05064	-4.609	4.06e-06	***
	Sector_Technology	-0.21481	0.05353	-4.013	6.00e-05	***
	Sector_Real_Estate	-0.09927	0.03719	-2.669	0.007601	**
	Sector_Financial_Services	-0.11369	0.05661	-2.008	0.044597	*