ECONOMICS AND BUSINESS ADMINISTRATION FACULTY
VILNIUS UNIVERSITY

FINANCE AND BANKING

Martynas Gintalas

MASTER THESIS

AKCIJŲ RINKOS REZULTATŲ IR FINANSINIŲ RODIKLIŲ SĄNTYKIS: MAŠININIO MOKYMOSI METODAS	**THE RELATIONSHIP BETWEEN STOCK PERFORMANCE AND FINANCIAL RATIOS: AN APPROACH BASED ON MACHINE LEARNING**

Master degree student _____

(signature)

Supervisor _____

(signature)

Supervisor Algimantas Laurinavičius

Date of submission of Master Thesis: 2022-01-10
Ref. No.

Vilnius, 2022

TABLE OF CONTENTS

The list of tables

The list of figures

INTRODUCTION

Financial markets can be considered the backbone of the global economy, offering liquidity and security from many types of uncertainties inherent in borrowing and saving. Financial performance, however, is highly impacted by the volume and accuracy of accessible financial sector data for an investor. Due to its importance, financial forecasting generates a great deal of study in a wide variety of interconnected areas of science, particularly recently in artificial intelligence's increasing usage in finance.

Lately, an exciting field of study has been evolving and is becoming more accessible to a wider variety of population. That is artificial intelligence, and research is focusing on its methods for processing randomness and non-linear relationships. Technological advancements and advances in computer and data science also increased public awareness of smart computing applications, thus stimulating the development of study in the area of machine learning. Financial market forecasting research employs a variety of machine learning algorithms, such as gradient boosting, random forests, and vector machines to forecast the values of stocks, yields and uncertainty of equity indexes, non-equity securities, and commodities (Henrique et al., 2019).

Stockholders or investors seek to build their fortune by purchasing businesses' shares, because they offer the greatest opportunity for long-term returns if risk-reward is considered. Although engaging in equity investment carries a higher risk, it often also carries a higher future profit. Generally, attractive investment prospects are identified by an examination of overall stock performance, as well as other financial indicators that indicates an investment's actual development over time. As an example, Anwaar (2016) argues that investors' perception of investment opportunities is shaped by the financial details of the firm. According to their source, the company's data is classified as internal or external. External knowledge is usually gathered from stock market, while internal data is collected from financial statements. As a result, investors will use the data to evaluate equity performance over time while making investment decisions. (Emamgholipour et al., 2013). Financial ratios are one of the products of financial data. They provide a quick snapshot of a company's performance. However, more than ever recently, fundamental financial information about a company may be insufficient or even misleading in predicting whether the company's share price will rise or fall. Therefore, this paper will employ machine learning algorithm to analyze the relationship between financial ratios and stock price performance of more than 4000 companies in the United States of America using financial data from 2014 to 2018.

The level of exploration of the topic

The previous scientific financial literature is full with controversy about the usefulness of financial ratios in predicting stock returns. A few decades of previous research have shown little

agreement about whichever of the ratios seems to have the highest correlation with stock returns. For example, Arkan (2016) claims that financial ratios were and are frequently used for forecasting purposes. Additionally, he asserts that they are useful measures of a financial and market success of certain company and may aid in forecasting potential performance. As per Lewellen (2004), stock market returns are predictable in general, as he discovered that perhaps the dividend yield ratio provides a strong predictive ability for stock returns.

Analogously, Öztürk's et. al. (2018) research discovered that earnings to price ratio and net profit margin are by far the most significant financial ratios for forecasting firm results. They assessed the relationship among financial ratios and firm results in the stock market using Two-way fixed effects models. Similarly, Musallam's (2018) findings contribute to the discussion, as he discovered that earnings per share and dividend yield have a strong correlation with stock market.

Previous research on the relationship between stock returns and financial ratios concentrated on older data and a relatively restricted number of firms, while also using comparably simplistic methodologies. As a result, extensive study covering as many companies as possible with the most recent financial details accessible is important. Then again, there is no general consensus in literature in establishing certain winners in correlation between financial ratios and stock price performance. Predicting a company's success is undeniably a difficult challenge, made much more difficult by using just a handful of financial ratios. Nevertheless, there is a strong need for decision-makers, such as financial researchers, lenders, fund managers, and shareholders, to define an appropriate and most current set of ratios for making reasonable forecasts.

Purpose of the study

The intention of this study is to examine the relationship between financial ratios and stock performance in the U.S. stock markets between the years of 2014 and 2018. It reviews the existing literature on financial ratios and their application to financial results prediction, as well as machine learning in the field of stock market prediction. Besides that, the aim is to develop and apply machine learning model for share price prediction that would incorporate a variety of financial ratios as parameters.

Research questions

By addressing the below research question and sub-question, the study aims at creating generalized statement regarding the relationship of related financial ratios and stock returns.

The research question, therefore, is following:

"What relationship exists between financial ratios and stock market performance in the United States of America from 2014 to 2018?"

Additionally, the sub-question is formed to determine if some financial ratios have a stronger link with stock returns than others:

"Which of the financial ratios possesses stronger relationship with stock price performance?"

Research Objectives

Conduct an examination of the research methodology and results of academic studies in the area of financial ratios and machine learning and their relationship;

Compile methodology for this paper and give an in-depth overview of how it is applied;

Perform the analysis using the chosen methods;

Explain the study's results and discuss its shortcomings, assumptions, and potential suggestions.

Research Methodology

The financial data is obtained using an API, from a third-party service that holds all publicly accessible and audited financial information for firms listed on the NYSE, NASDAQ, AMEX, and other exchanges. Python is used to clean and manipulate the data. Using stratified sampling approach, the dataset was split between training and test sets. Explanatory data analysis, linear regression model, and machine learning algorithm library Extreme Gradient Boosting (XGBoost) were used in order to assess the relationship between financial ratios and stock returns. The latter methods helped to calculate feature importance scores for financial ratios and to conduct Shapley Additive Explanations Value (SHAP) analysis.

Structure of the Research

This research combines artificial intelligence and corporate finance principles while maintaining a quantitative focus on the topic. The theoretical structure of the study is shown in Figure 1, which is divided into the thesis's major research topics and theoretical foundations. The thesis' theoretical foundations consist of financial ratio analysis which is a subset of fundamental analysis and also regression analysis which comes from supervised machine learning.

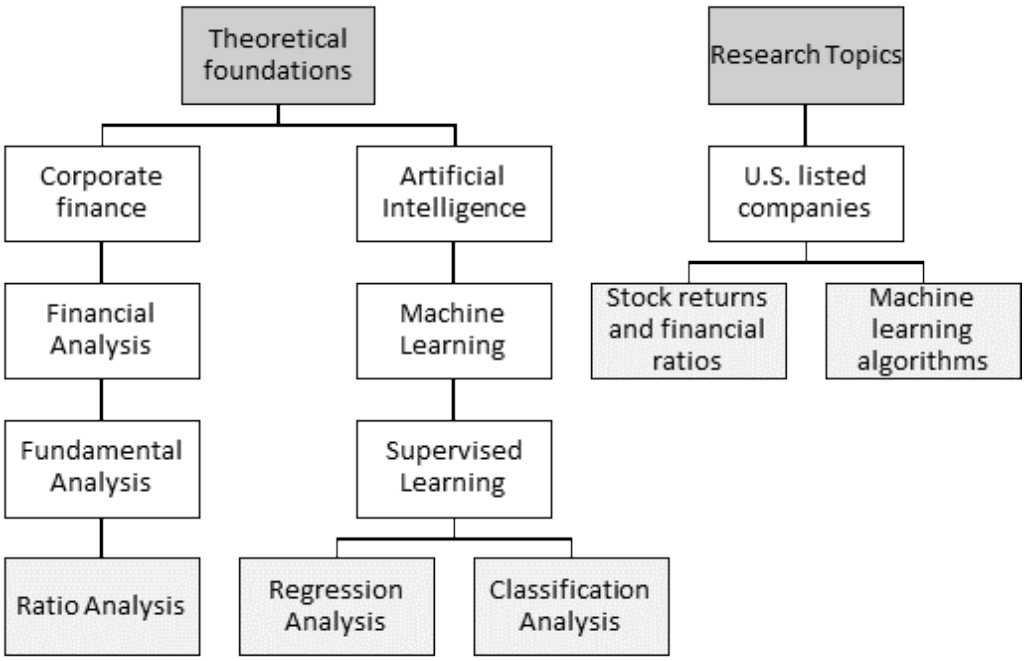


Figure 1. Research structure.

The introductory chapter establishes the motivation by defining the objective, aim, purpose, and focus. Additionally, it provides a theoretical basis. The first chapter then presents findings on conducted study of the literature on the use of financial ratios. Additionally, the fundamentals of machine learning are discussed, as well as machine learning in the field of stock market prediction. The second chapter discusses the research methodology used and the workflow associated with it. This section contains a summary of the data and technology that were used. Additionally, the fundamental principles, tools, and techniques for developing the model are explored, including techniques for data processing and assessment, as well as concepts for various machine learning models and the principles of feature importance and SHAP values. The third chapter documents, presents, and interprets the findings.

1. THEORETICAL ANALYSIS OF FINANCIAL RATIOS AND MACHINE LEARNING

1.1. Predicting Financial Performance Using Financial Ratios

This paper is centered on the principles of the relationship between financial ratios and stock market returns. Profitability ratios acting as indicators of company success and returns, and mathematical and machine learning-based prediction methods. A review of the literature seeks to assess which financial ratios have a strong correlation with stock returns, regardless of whether mathematical or machine learning approaches are used. Additionally, prior research data and methodologies are explored.

Accounting techniques are perhaps the most often used tool for evaluating business results and providing critical internal statistics on a company's valuation and market capitalization. Accounting procedures and their implementation have been widely used for years, especially in forecasting future success of businesses. Additionally, they can be used to forecast stock market performance. Additionally, financial ratios are often used to determine a company's ability to repay debts (Lewellen, 2004).

It has been noticed quite a while ago that simple financial ratios can provide a lot of information about a company’s state. For example, Beaver (1968) wrote in his paper already in 1968 that empirical evidence shows that financial ratios predict potential insolvency and insolvency vulnerability for at least five years prior to a firm's collapse. Firms default rarely, and it can be very costly to their providers of finance and reorganization or liquidation can be very painful. The argument that investors use ratios in determining solvency is consistent with the idea that investors predict loss earlier than ratios. This finding is further confirmed by the probabilistic tables and the relation between ratio and return forecasts. Given the lack of ideal correlation between the estimates, investors either relied on non-ratio sources of knowledge, or they misused the ratios, or both.

In a more recent study Martani (2009) pointed out that based on the empirical results, it can be inferred that financial ratios, company scale, and cash flow from operations influence return on investment. Variables that show a major impact on market adjusted return include profitability ratios, such as net profit margin (NPM), return on equity (ROE), or price to book value (PBV), and Net sales divided by Total Assets (TATO). In his conclusion the author mentions that ratios are valuable to investors and they help in making decisions on investments.

A simple glimpse into the scientific literature reveals the widespread usage of financial ratios. Many articles have been written on the relationship between a firm’s performance, defaults, bankruptcies and financial ratios. During this study the search was tailored to articles that

discussed relationship between the ratios and stock returns. These papers often differ from others based on their peculiar methodologies and datasets used. As independent variables, numerous variations of financial ratios were used, as well as various mathematical or machine learning-based methods.

According to financial theory, financial ratios may provide visibility into a company's financial results through rearrangement of the financial statements: the firm's efficiency and value, the potential to allocate cash flow, short-term liquidity, and the size of debt funding. Additionally, financial ratios are used to forecast potential results. They are used as parameter estimates in statistical modeling in order to foresee financial instability, defaults, and bankruptcies for example (Ross et al., 2002).

Equally, financial ratios are used to model associations between financial ratios and market returns. This is accomplished by the use of a more or less unique set of financial variables or by the use of novel mathematical or machine learning techniques (Chairakwattana & Nathaphan, 2014) Although the majority of the examined published papers succeed in modeling the association between financial ratios and stock performance, they fall short of identifying and emphasizing the main characteristics that are important for evaluating firm results. The following table summarizes prior analysis on the relationship between stock returns and financial ratios.

Table 1. Past works.

Author	**Sample/data**	**Location**	**Period used**	**Technique**	**Significant ratios**
Anwaar (2016)	FTSE-100	UK	2005-2014	Panel regression	EPS; Net Margin and ROA
Arkan (2016)	15 firms	Kuwait	2005-2014	Multiple regression model	ROA, ROE, P/E, EPS
Chairakwattana, Nathaphan (2014)	395 firms	Thailand	2001-2011	Bayesian model averaging	Book to market
Chen, Shen (2009)	483 obs.	USA	1961-2001	Markov-switching model	P/E, P/D
Delen et al. (2013)	2345 obs.	Turkey	2005-2011	Decision trees, CHAID, C&RT	Earnings before tax to equity, profit margin
Emamgholipour et al. (2013)	80 firms	Iran	2006-2010	Multiple regression models	EPS, PE, book to market
Fama, French (1988)	NYSE	USA	1941-1986	WLS (Weighted least squares)	Dividend yield
Jermsittiparsert et al. (2019)	10 firms	ASEAN	2012-2016	Multiple regression	Assets' turnover, assets growth, quick ratio
Kheradyar et al. (2011)	960 firms	Malaysia	2000-2009	GLS (Generalized least squares)	Earnings yield, dividend yield, book to market

Lewellen (2004)	NYSE	USA	1946-2000	OLS (Ordinary least squares)	Dividend yield, book to market, E/P
Martani et al. (2009)	JSX	Indonesia	2003-2006	Regression analysis	NPM, ROE, TATO, PBV
Mussallam (2018)	26 firms	Qatar	2009-2015	WLS & OLS	EPS, earnings yield, dividend yield
Öztürk & Karabulut (2018)	14 firms	Turkey	2008-2016	Two-way fixed effects models	Earnings to price, profit margin
Pech et al. (2015)	29 firms	Mexico	1995-2011	Panel regression	EPS, P/E, firm value to EBITDA, DY, sales growth
Petcharabul & Romprasert (2014)	22 firms	Thailand	1997-2011	OLS (Ordinary least squares)	ROE and P/E

However, as Delen et al. (2013) pointed out, there are significant differences in the financial ratios chosen, the methodology used to calculate them, and other financial ratios used in previous studies. For example, Delen's (2013) study used thirty-one financial ratios to assess firm performance and used Decision tree - a machine learning algorithm, while Ma et al. (2018) analyzed 25 financial ratios, Wang et al. (2009) reviewed 50 ratios, and Musallam (2018) examined 9 financial ratios.

Numerous academic articles conducted in prestigious papers, on the other hand, used economic factors to enhance financial ratios. Chen & Shen (2009), for example, used 6 economic and financial factors, Chairakwattana & Nathaphan (2014), and Tsai et al. (2011) used 12 economic metrics and 19 ratios. There is no consensus in the literature about the optimal set of variables to use to accurately predict a company's results. Pech et al. (2015) conclude that the most frequently used financial ratio categories for evaluating company success are profitability, leverage, and margins and that the most frequently used five ratios are dividend yield, price to earnings, earnings per share, firm value to EBITDA and revenue growth. Although these ratios are widely used in real life practice, there is some variance within the research community.

Delen et al. (2013), for example, used asset structure ratios, liquidity, growth, turnover, solvency and profitability ratios. Their paper analyzed profitability using seven ratios: operating expense-to-net revenue, return on assets, profits before taxes-to-equity, EBITDA margin, net profit margin, gross profit margin, return on equity. Ma et al. (2018) analyzed risk, liquidity, solvency, growth capability, operational capacity and profitability financial ratio categories. Earnings per share, return on income, net profit rate, book value per share, return on equity and gross profit rate are six profitability measures. Musallam (2018) considered return on assets ratios, dividend yield, earnings yield, price to earnings, earnings per share, return on equity and stock to book value ratios, as well as net profit margins and dividend earnings. Chairakwattana & Nathaphan (2014) examined only financial factors such as dividend yield, book-to-market ratio

and earnings yield, while Petcharabul & Romprasert (2014) have used current ratio, inventory turnover, debt-to-equity ratio, price-to-earnings ratio and return on equity. Anwaar's (2016) analysis makes use of earnings per share, fast ratio, return on equity, return on assets, and net margin.

Chen and Shen (2009) utilized economic and financial factors, such as the price to earnings ratio, price to dividend ratio, and yield spread to forecast stock returns in the United States. Tsai et al. (2011) were using the profitability ratios, which were: EBIT, net assets per stock, return on investment and return on equity in their report. According to Kheradyar's et al. (2011) study, the dividend yield, earnings yield, and book-to-market ratio all have an effect on stock market performance. Wang et al. (2009) forecasted potential earnings per share adjustments by examining six profitability ratios: return on equity, equity to capital assets, return on total assets, gross margin, and net margin. Emamgholipour et al. (2013) estimated company success using the earnings per share ratio, price to sales, and market valuation to book value of shares. Subsequently, Oztürk and Karabulut (2018) examined the relationship between stock returns and financial ratios using current ratio, earnings-to-price ratio, and net margin.

Fama and French (1988) demonstrated that dividend yield could be used to forecast portfolio returns. Additionally, they said that operating revenues can be preferred to net income due to its noisiness. Lewellen (2004) illustrated that the book-to-market, earnings-price, and dividend yield ratios could all be employed to estimate portfolio returns. According to Delen et al. (2013), the EBT-to-equity ratio and net margin are the critical ratios for forecasting business success as measured by ROA and ROE. Musallam (2018) discovered a strong beneficial relationship between equity returns and dividend rate, earnings per share and earnings yield. Chairakwattana & Nathaphan's (2014) research discovered that the book to business ratio is a critical variable in predicting stock returns. Petcharabul & Romprasert (2014) discovered two significant findings. To begin, it was discovered that the return on equity and price to equity have a substantial correlation with returns at the 95% confidence mark. Second, product volatility, debt-to-equity, and current ratios have no relationship with returns. Pech et al. (2015) suggest that 1-year stock returns can be predicted by revenue rise. On more conventional measures such as price to earnings, dividend yield, earnings per share and company valuation to EBITDA a comparable argument was made. Anwaar's (2016) research discovered a negative correlation between returns and earnings per share. Additionally, Anwaar (2016) discovered a strong correlation between the yield on investment and net margin ratios and stock market returns. Chen and Shen (2009) observed that the price to earnings ratio and dividends are related to equity returns. As per Kheradyar et al. (2011), high predictive capacity can be found in book-to-market ratio, as opposed to earnings yield and dividend yield. Besides this, they conclude that combining the ratios

improves stock return forecasting since every ratio is distinct but they complement each other. Emamgholipour et al. (2013) identified a positive correlation between earnings per share and current-year portfolio returns. However, the price to earnings and stock valuation to book value have an adverse association for returns in the current and potential years. Profit margins and earnings-to-price ratios are positively correlated with returns and typically deliver higher returns over the resulting forecast period (Oztürk & Karabulut, 2018). Kim and Upneja (2014) learned that a minimal debt-to-equity ratio and strong growth rates in profits and assets decrease the likelihood of instability and loss, but doesn't mention the profitability ratios. Wang et al. (2009), Sun et al. (2011), and Tsai et al. (2011) doesn't rate the financial ratios' significance at all.

Previous researchers have presented a comprehensive variety of financial ratios for stock return estimation, which can be analyzed by the samples and varied methodologies utilized by scholars. Then again, the study of various markets can help justify the heterogeneity. Like Ma et al. (2018) noted, while transitioning from one economy to the next, as is the case in several East Asian countries, whereby efficient market theory doesn't automatically incorporate into financial data, fundamental analysis-based forecasts may be applied effectively. Nevertheless, numerous capital markets and time ranges are covered, ranging from emerging to mature economies.

For example, Anwaar (2016) conducted a study from 2005 to 2014 on the listed firms in FTSE-100 Index on the London Stock Exchange. Also, two Turkish studies were chosen for this paper. Delen et al. (2013) examined publicly listed firms on Borsa Istanbul exchange between 2005 and 2011, while Oztürk & Karabulut (2018) investigated a dataset of 14 firms and 448 findings from the ICT market mentioned on the Borsa Istanbul between 2008 and 2016. Additionally, some articles examined the financial market in North America. Fama and French (1988), in particular, examined NYSE returns through 1941 to 1986. Lewellen (2004), similarly, used data from the NYSE index from 1946 to 2000. Kim and Upneja (2014) used a dataset of restaurants and food service sectors in NA during the years 1988 to 2010. Chen and Shen (2009) conducted a study from 1961 to 2001 on US bear financial markets. Pech et al. (2015) examined the 1995–2011 performance of 29 Mexican firms. Zieba et al. (2016) analyzed data from the EMIS index, which covers Polish firms from 2000 to 2013.

A sizable proportion of related studies is discovered to have been performed in Asian countries, with the majority concentrated in China. Ma et al. (2018), for example, used data from 60 publicly traded companies spanning 2011 to 2015. Sun's et al. (2011) study consisted of 692 Chinese firms listed on the Shenzhen and Shanghai stock exchanges between 2000 and 2008. Wang et al. (2009) have used 3181 findings from either the Shenzhen or Shanghai markets between the years 1996-2005. Additionally, Tsai et al. (2011) used Taiwanese semiconductor industry sales statistics from 2002 to 2006, based on quarter performance. Numerous examinations

were also conducted in Middle East and Southeast Asia. Musallam (2018), for example, analyzed the market performance of 26 Qatari public firms throughout 2009-2015. Emamgholipour et al. (2013) examined 80 publicly traded companies in Tehran between 2006 and 2010. Kheradyar et al. (2011) used a dataset of 960 companies listed on Bursa Malaysia during 2000-2009. Furthermore, two tests were repeated in Thailand. Chairakwattana & Nathaphan (2014) analyzed data from listed Thai businesses during 2001-2011, whereas Romprasert and Petcharabul (2014) analyzed quarterly investment performance of 22 technology companies listed on the Thai Stock Exchange during years 1997-2011.

Statistical approaches have been extensively used in previous research, despite the fact that these methods often employ assumptions about linearity and normality when it comes to financial information. However, they have been included in a number of tests. For instance, Lewellen (2004) estimated a linear regression model using the ordinary least squares (OLS) procedure. Likewise, Petcharabul & Romprasert (2014) examined the relationship between financial ratios and returns using OLS. Musallam (2018) analyzed data using ordinary least squares (OLS) and weighted least squares (WLS). Kheradyar et al. (2011) estimated regression predictions using generalized least squares (GLS). Anwaar (2016) and Pech et al. (2015) utilized panel regression analysis, whilst Emamgholipour et al. (2013) used various regression models and panel data econometrics to evaluate their theories. Chairakwattana and Nathaphan (2014) made a decision to use averaging Bayesian models. Ma et al. (2018) analyzed correlations using three different techniques. Chen and Shen's (2009) study examined four distinct Markov-Switching styles. Oztürk & Karabulut (2018) examined their framework accuracy using two-way fixed effects simulations.

Furthermore, there is machine learning, and its approaches that have been implemented to the same subject in the literature. Machine learning's dominance over statistical methods can be demonstrated by the fact that it is not as assumption-dependent and, in most situations, avoids them. Additionally, machine learning extracts trends from data with fewer human intervention. The decision tree algorithm family, a well-known machine learning algorithm, is heavily discussed in the literature. They are deemed readily understandable, and visualizing the findings is beneficial. Delen et al. (2013), for instance, used four algorithms coming from a decision tree type of machine learning.

Comparably, Wang et al. (2009) adapted boosting and bagging, which in simpler terms is merging similar type of predictions, to three models of decision trees. Kim and Upneja (2014) used cross-validation to determine the quality of prediction produced by AdaBoost and decision tree algorithms. Tsai et al. (2011) tested a group of multi-layer perceptron, which in other words is neural network, logistic regression models and decision trees (CART) utilizing 5-fold cross-

validation. While Sun et al. (2011) incorporated two AdaBoost algorithms and a Support Vector Machine into their study. Zieba et al. (2016) contrasted sixteen approaches and discovered that excessive gradient boosting outperformed random forest, decision tree, AdaBoost, and support vector machine.

That being said, there is a relatively recent algorithm named Extreme Gradient Boosting (XGBoost) that hasn't yet been often used in academic research, especially on the financial ratio and stock return relationship. This can simply be seen by number of papers that have used this algorithm. For example, before 2017 there were less than 100 studies that employed this machine learning method, and only in the year 2020 it became slightly more popular as academic researchers started to notice it. Nonetheless, it is rapidly gaining popularity. Due to its experiment-tested performance on large-scale issues, XGBoost is becoming a commonly deployed and extremely common platform among Kaggle competitors and Data Scientists. Zhang and Chen (2019) developed a multi-factor stock picking model using Extreme Gradient Boosting (XGBoost) to outperform the benchmark in their paper. They used 62 accounting, pricing, emotional, and technical metrics as features for the XGBoost classifier, with the aim of identifying stocks in the CSI 300 Index that would be expected to surpass the sector by extraordinary returns. Over a 32-month span, the average return on the similarly weighted stocks chosen by XGBoost was 134 percent, far greater compared to benchmark return of 28 percent. Simultaneously, the authors contrasted the XGBoost algorithm to two other classifiers, Support Vector Machines (SVM) and Logistic Regression (LR) and discovered that the XGBoost algorithm outperformed the other two algorithms in terms of AUC (Area under the ROC Curve) and final cumulative returns, showing that the XGBoost algorithm is suitable for dynamic nonlinear equity markets.

1.2. Prediction of the Stock Market Through Machine Learning

Humans solve challenges by amassing a vast volume of present knowledge, a process known as learning. Machine learning operates similarly, albeit on a more primitive level, at least so far. For instance, Sebastian Thrun (2019), a computer scientist and educator from Germany says that "ML is distinct from other types of computer programming in that it transforms the inputs of an algorithm into outputs using statistical, data-driven rules that are automatically derived from a large set of examples, rather than being explicitly specified by humans". More precise definition was provided by Mitchell (2006): "…machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc."

Thus, the primary focus of attention in the machine learning area is on the development of algorithms and the refinement of previously developed algorithms using provided data. Additionally, it offers pertinent information for the project. If the data has been analyzed and presented, a machine learning algorithm may determine how to draw conclusions by making generalizations from the data. Thus, the emphasis is on the algorithms that are responsible for finding the optimal mathematical function for producing a meaningful result provided the inputs. Machine learning aims to make forecasts that have a reasonable error margin. In machine learning, intellect is primarily in the recognition of features, and an algorithm is simply learning to associate these features in order to draw appropriate conclusions (Ketkar, 2017).

Machine learning can be classified into three categories of learning: unsupervised learning, reinforcement learning, and supervised learning. This paper uses supervised learning; therefore, it is the most relevant here. Its aim is to discover the mathematical formula that would best approximate the relationship between the inputs and eventually provide results. Additionally, supervised learning is subdivided in regression and classification analysis scientific fields. As per Sarker (2019) “The goal of classification is to accurately classify the activity class labels of instances whose contextual features or attribute values are known, but class values are unknown.” While in regression analysis, the algorithm's job is to generate accurate forecasts by mining the necessary details from the data.

Classification is the process of grouping data. The main process of it is to determine to which of a collection of groups a new data point belongs to. This is done by using a training set of data comprising known-category observations. Numerous practical situations may be constructed as classification issues, for example classifying an email as “spam” or “non-spam” (Tang, Alelyani, & Liu, 2014). This can be accomplished in two stages using the grouping technique. To begin, the method constructs a model, which is used for the class characteristic, as a feature of the dataset's other variables. Following that, it applies a beforehand constructed model to the newly created and previously unknown datasets in order to determine the relevant class of each record (Fernández-Gavilanes et al., 2016, Heydari et al., 2016).

The aim of regression analysis is to identify relationships in the input datapoints that would allow for the development of accurate outputs or forecasts. Forecasting stock returns is a form of regression analysis procedure that uses financial ratios from a stock as input data. The dataset is generally separated into two sections: the training set and the test set. By learning from labelled results, the training set is commonly employed to estimate the parameters of the model and model itself. Additionally, the test set is being used to evaluate the model's performance on previously unseen results (Singh et al., 2010).

According to Atsalakis & Valavanis (2009), machine learning techniques have been thoroughly studied and applied to stock market analysis. As a result, the simplest machine learning algorithm would be ordinary least squares or linear regression. It is analogous to statistical linear regression in that the algorithm fits a linear function to the input data and makes predictions with a small margin of error, but also differs, because, as opposed to statistical method, in machine learning the machine is learning from the past. Moreover, because the parameters are constant, linear regression may be considered a parametric procedure. Additionally, the function's design associating input and output data is already presumed.

In comparison, no structure is defined for non-parametric algorithms. Other than that, the structure of the model is dictated entirely by the input data, making it more adaptable to the nature of the training set distribution. Nearest neighbors are a common example of this form of algorithm, as well as algorithms based on decision trees such as gradient boosting and random forest, which uses a large number of decision tree-based models. This style of algorithm is an excellent place to start for the modeling phase since it has a high tolerance for unstructured data and a high efficiency.

Henrique et al. (2019) and Atsalakis & Valavanis (2009) conducted analyses of recent articles on equity market prediction using both stock indices and stocks itself. Although Atsalakis & Valavanis (2009) analyzed over a hundred published papers categorizing them as well-developed or developing markets, Henrique et al. (2019) examined 57 articles mostly focusing on Taiwanese markets and North America. According to Atsalakis & Valavanis (2009), the S&P 500 index is the most frequently forecasted index, whereas the Singapore exchange would be the most frequently predicted emerging market. Although the subject is extensively researched in China, just a few studies utilized Chinese stock market information (Henrique et al., 2019).

Numerous economic forces, such as market dynamics, have an effect on financial markets. As a result, consumer behavior is widely analyzed in order to improve the reliability of financial market predictions. Market analysis is commonly classified into basic and scientific study methods. Both methods, in their own special forms, aim to determine market trends and forecast potential asset movements (Cavalcante et al., 2016). Although there is no universal consensus about which method is the most effective for market analysis, many studies have demonstrated which methodology is the most common. For example, Henrique et al. (2019) discovered that indicators of technical analysis are by far the most common input variables, followed by fundamental analysis variables. They discovered that it is uncommon to use all types of metrics concurrently.

Equally, Cavalcante et al. (2016) assert that technical research actually utilizes business dynamics and as such is a generally more applied method. Atsalakis & Valavanis (2009)

discovered that technological research metrics, which usually consist of between two and twenty-five variables, are sometimes paired with regular or historical prices. Although these two methods have been extensively researched and implemented to the prediction field, Cavalcante et al. (2016) note that data mining is often used to derive valuable knowledge from primary data in certain primary studies. In conclusion these authors state that variable counts varied significantly through studies. As per Atsalakis & Valavanis (2009), the average number of variables was around seven, although some researchers also used as much as two or nearly sixty. The most highly valued technical factors are those that relate to asset values: a stock's opening and closing prices (Cavalcante et al., 2016).

Forecasting equity market movements is a well-studied topic. Numerous novel methods are discussed, as well as a broad variety of applied algorithms. Patel et al. (2015), for example, evaluated forecasts using a variety of standard algorithms, including support vector machines and random forest. When features are treated as trend deterministic data, the random forest outperforms other models if viewed on average. Imandoust and Bolandraftar (2014) evaluated the efficiency of few classification models designed to forecast the trajectory of a stock index's movement. They used random forest, decision tree, and naive-Bayesian classifier frameworks to isolate and combine fundamental and technical measures. Both models generated adequate predictions, however the decision tree surpassed both other algorithms.

Decision tree algorithms were widely used to estimate a firm's profitability success (Wang et al., 2009) as well as to foresee economic difficulties (Kim & Upneja, 2014). Additionally, ensemble models dependent on decision trees, such as Random Forest, AdaBoost and Gradient Boosting, are used to forecast depression and financial difficulties (Zieba et al., 2016). Similarly, Sun et al. (2011) used a support vector machine to measure the financial pressure of Chinese firms using a variety of profitability ratios.

2. METHODOLOGY FOR RESEARCHING THE RELATIONSHIP BETWEEN STOCK PERFORMANCE AND FINANCIAL RATIOS USING MACHINE LEARNING

2.1. Aim, purpose, model and research questions of the research

The aim of this paper is to inspect the relationship between financial ratios and stock returns in the U.S. stock exchanges from 2014 to 2018 using novel machine learning technique and a large dataset. Quantitative techniques were employed to determine which financial ratios have a strong correlation with stock return. Additionally, a machine learning model was used to investigate which ratios could forecast stock returns, utilizing a set of financial ratios that have been shown in the literature to have a strong correlation with company share price results.

This research adds to the existing literature by using a relatively great size and current dataset from U.S. stock exchanges, as well as employing a novel machine learning technique. To the novelty of this study also contributes the fact that not only regression analysis was used, but also machine learning-based classification analysis. Most of the prior studies were employing regression analyses. As a result, the model used in this analysis produces unique prediction accuracy outcomes. The discovery of strongly associating financial ratios, on the other hand, is comparable to past scientific work. Given that predicting business success is a critical component of the financial industry work, it will be beneficial to understand which ratios have the strongest correlation with stock returns.

By addressing the below research question and sub-question, the study aims at creating generalized statement regarding the relationship of related financial ratios and stock returns.

The research question, therefore, is following:

“What relationship exists between financial ratios and stock market performance in the United States from 2014 to 2018?”

Additionally, the sub-question is formed to determine if some financial ratios have a stronger link with stock returns than others:

“Which of the financial ratios possesses stronger relationship with stock price performance?”

The literature review addresses the primary research issue in part by elucidating the appropriate collection of financial ratios to use in the modeling. The sub-questions are formulated to aid in the resolution of the primary question. To address these problems, the financial statements of more than 4000 publicly traded firms in U.S. were examined.

2.2. Workflow of the study

The research takes a quantitative scientific approach with the aim of producing generalizations about the relationship between related financial ratios and stock return expectation. To address the thesis study concerns, the financial statements of publicly traded firms in U.S. was analyzed. The primary steps taken when examining the relationship between financial ratios and stock returns using sophisticated machine learning methods are depicted graphically in a Figure below. Historically, financial success of businesses was measured using conventional statistical approaches - GLS and OLS. Although the approach used in this study has not been widely adopted in the field of quantitative stock return modeling, it has picked up traction in recent studies using decision tree models, for example.

The very first phase was data collection, which included obtaining raw data from a third-party service. Followed by workspace creation in Python and developing the dataset. All the empirical portion of the study was done using Python. Next step was exploratory data analysis in order to understand it and prepare for cleaning procedures. After a clear picture of dataset, cleaning was done, where missing values and outliers were handled accordingly. Afterwards, visualizations of relationships and distributions, feature engineering and selection was performed. Finally, prediction modelling using regression and classification methods was performed by employing machine learning algorithm XGBoost.

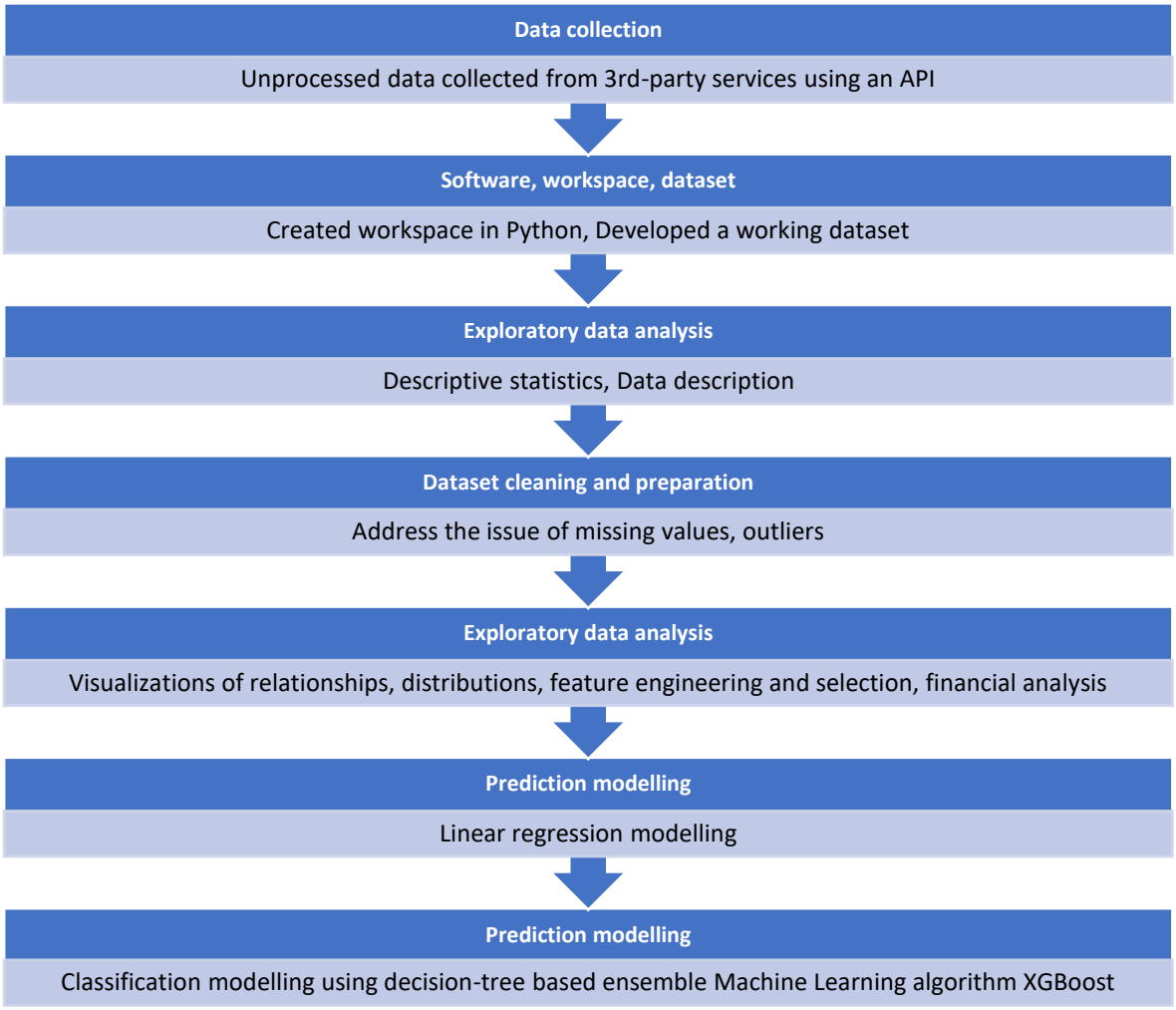


Figure 2. Workflow of the paper.

2.3. Software and data preparation

This section discusses the data and applications that were used for modelling. The study was performed using Python 3.8 programming software, with the help of following packages and libraries: for scraping and reading data Pandas Datareader, Json, Urllib, for data exploration and manipulation Pandas, Numpy, for visualizations and plotting Matplotlib, Seaborn, Plotly, Cufflinks, for predictive data analysis and machine learning Sklearn, to visually explain machine learning – Shap, Eli5, and for machine learning algorithm Extreme gradient boosting (XGBoost) model was implemented.

Table 2. Used tools.

Name	Address	Type
Python 3.8	https://docs.python.org/3.8/	Programming language
Pandas Datareader	https://pandas-datareader.readthedocs.io/en/latest/	Scraping, reading data
Json	https://docs.python.org/3/library/json.html	Scraping, reading data
Urllib	https://docs.python.org/3/library/urllib.html	Scraping, reading data
Pandas	https://pandas.pydata.org/	Data exploration, manipulation
Numpy	https://numpy.org/	Data exploration, manipulation
Matplotlib	https://matplotlib.org/	Visualizations
Seaborn	https://seaborn.pydata.org/	Visualizations
Plotly	https://plotly.com/	Visualizations
Cufflinks	https://github.com/santosjorge/cufflinks	Visualizations
Scikit Learn	https://scikit-learn.org/stable/	Predictions
SHAP	https://shap.readthedocs.io/en/latest/index.html	Visual explanations of ML
Eli5	https://eli5.readthedocs.io/en/latest/overview.html	Visual explanations of ML
XGBoost	https://xgboost.readthedocs.io/en/latest/	Machine learning library

Audited financial statements for years 2014-2018 were scraped from third-party service Financial Modeling Prep (FMP). It was done through their Application Programming Interface (API) using Python. The data set included all publicly traded stocks in major U.S. stock exchanges, which are NYSE, NASDAQ, AMEX. For predictive modelling purposes stock price variance was needed, which was scraped from Yahoo finance. However, as Yahoo closed down their API, it became problematic to access stock price variance information when it comes to a large number of companies and latest data. For classification purposes a class column was added. Each company got assigned either 1 or 0. If stock price variance was positive throughout the year the company was assigned 1, if price variance was negative the class that stock got assigned was 0. Data was obtained and cleaned year by year in different files, and only for modelling and predictive purposes it was joined into one dataset. For research and presentation purposes, following tables and figures will be from year 2018 dataset unless noted otherwise. Figure 3 provides basic information about the dataset from year 2018.

```
<class 'pandas.core.frame.DataFrame'>
Index: 4392 entries, CMCSA to ZYME
Columns: 224 entries, Revenue to Class
dtypes: float64(222), int64(1), object(1)
memory usage: 7.5+ MB
```

Figure 3. Data set description.

Looking at the Figure 3 we can see that 2018 dataset has 4392 companies, 224 columns, which are various financial data points. Moreover, that 222 columns are numeric, 1 integer, which is the class column, and 1 object column, which is categorical and simply describes company's sector.

Figure 4 shows summary statistics and data structure of the untreated dataset for several random variables.

	Revenue	Revenue Growth	Cost of Revenue	Gross Profit	R&D Expenses	SG&A Expense	Operating Expenses	Operating Income	Interest Expense	Earnings before Tax	3Y Dividend per Share Growth (per Share)	Receivables growth	Inventory Growth	Asset Growth	Book Value per Share Growth	Debt Growth	R&D Expense Growth	SG&A Expenses Growth	2019 PRICE VAR [%]	Class
count	4.35E+03	4253	4.21E+03	4.33E+03	4.16E+03	4.23E+03	4.21E+03	4.36E+03	4.21E+03	4.32E+03	4067	4268	4160	4178	4121	4128	4133	4144	4392	4392
mean	5.12E+09	3.455278	3.14E+09	2.04E+09	1.18E+08	9.01E+08	1.44E+09	6.54E+08	1.00E+08	5.58E+08	0.006081	36.768524	0.183066	1.389013	0.26253	9.928446	0.091891	0.15361	20.803948	0.693534
std	2.05E+10	195.50491	1.51E+10	7.68E+09	9.33E+08	3.66E+09	5.53E+09	2.97E+09	3.78E+08	2.64E+09	0.239653	2347.0792	4.688013	35.123904	5.612666	363.71773	0.823281	0.839647	82.622147	0.461078
min	-6.89E+07	-3.4615	-2.67E+09	-1.82E+09	-1.04E+08	-1.40E+08	-4.28E+09	-1.46E+10	-1.41E+09	-2.18E+10	-1	-1	-1	-0.9991	-32.2581	-1	-1	-1	-99.864779	0
25%	6.50E+07	0	3.42E+06	3.62E+07	0.00E+00	2.06E+07	4.22E+07	-5.51E+06	0.00E+00	-1.00E+07	0	-0.048075	0	-0.0367	-0.1086	-0.08285	0	-0.00465	-7.477173	0
50%	4.98E+08	0.0749	1.74E+08	2.22E+08	0.00E+00	9.39E+07	1.81E+08	4.20E+07	5.69E+06	2.73E+07	0	0.0102	0	0.03475	0.0261	0	0	0.0657	17.639393	1
75%	2.46E+09	0.1885	1.30E+09	9.77E+08	1.45E+07	4.12E+08	6.80E+08	2.86E+08	5.82E+07	2.24E+08	0.04205	0.1859	0.08005	0.160575	0.1384	0.115425	0.0097	0.167625	39.625879	1
max	5.00E+11	12739	3.73E+11	1.27E+11	2.88E+10	1.07E+11	1.07E+11	7.09E+10	9.17E+09	7.29E+10	4.0791	153332.33	293.473	1184.9938	313.3958	17646.824	36.8981	43.7188	3756.7163	1

Figure 4. Summary statistics.

In order to obtain accurate forecasts in the area of machine learning, it really is critical to perform adequate sampling and data transformation (Géron, 2017). Following information refers to data cleaning and transformation.

Figure 5 provides information about the distribution of the classes, where, as stated earlier, if stock price variance was positive throughout the year the company was assigned 1, if - negative the class was 0. Figure 6 shows distribution of the variable “Sector”, which is a categorical type of variable.

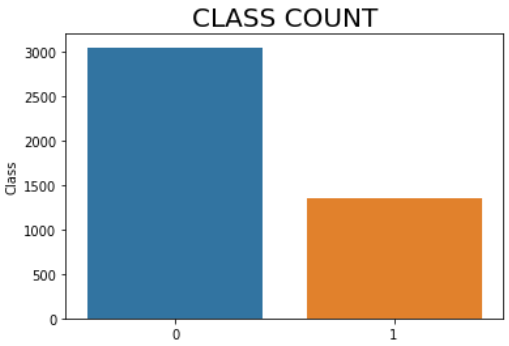


Figure 5. Class count.

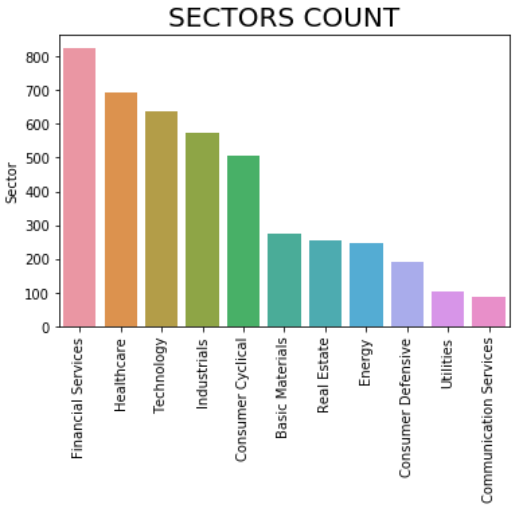


Figure 6. Sectors’ count.

Figure 5 reveals that the dataset is not balanced in terms of class, which should be kept in mind for later when data is split into training and test sets. While from Figure 6 we can see that we have 11 sectors and at least 5 them have more than 500 companies, while the other 6 contain less than 300. This must be considered for future use in machine learning algorithms: too small number of samples could cause overfitting and similar inconveniences (Géron, 2017). Going further, it is important to ensure that the target data is logical, or in other words makes sense. To be more specific, the price variation column is of major significance and has to be checked for

mistakes, which can be mistyping or just unreasonable values. It can also be described as searching for unusual drops or peaks, which would indicate that the stock decreased or increased by an extreme amount compared to sector's overall pattern (VanderPlas, 2019).

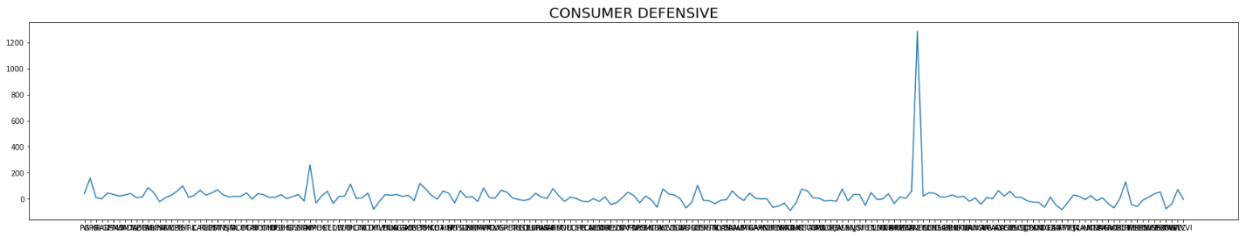


Figure 7. Consumer defensive outliers.

From figure 7 it can be interpreted that there are some unusual peaks in Consumer Defensive sector. Other sectors can be found in appendix: Consumer Cyclical, Technology, Industrials, Basic Materials, and Healthcare sectors seemed to have unusual peaks. This implies that for unknown to us reasons some stocks had unusual price variances. The next step would be to find out if these price growths were organic, in other words based on reasonable trading activity (VanderPlas, 2019). Based on the previous figures we can do that by searching for stocks, which experienced price increases for more than 500%. One of such companies was with a trading ticker AXSM.

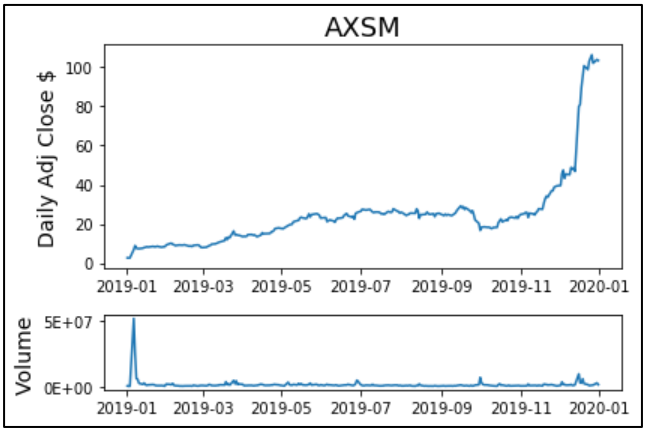


Figure 8. AXSM unusual gain.

It can be seen from Figure 8 that the Volume, which is trading activity, does not correspond with late price increase. Six such stocks were identified in 2018 dataset, and for research purposes were dropped from dataset. If left untreated, these data points can skew the distribution and, in that way, make it hard to realize the true picture of whole data set (VanderPlas, 2019).

Figure 9 shows Consumer Defensive sector after elimination of these stocks. It can be seen that the sector's distribution of price variance looks reasonable. Figures for all other sectors can be found in appendix.

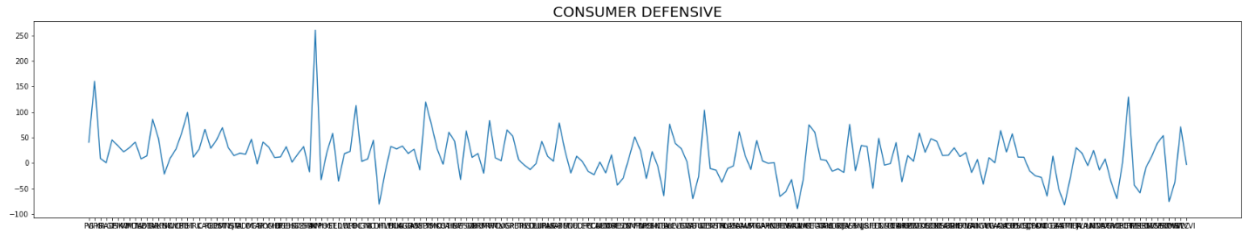


Figure 9. Consumer defensive after cleaning data.

Next step in data clean-up can be to check for missing values (NaN) and 0 values (VanderPlas, 2019). Figures 10 represents the NAN values and 0 values throughout the whole dataset.

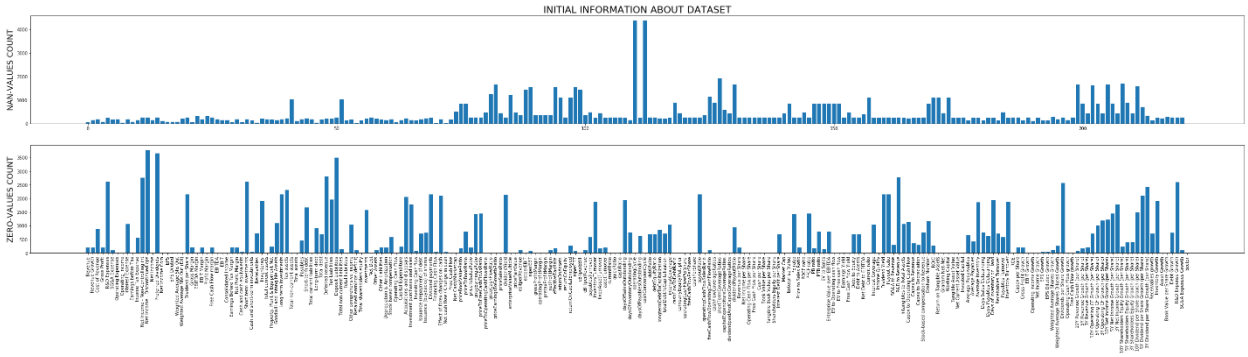


Figure 10. NAN and 0 values.

It can be seen that there are quite a lot of NAN and 0 values. To get a better picture of the situation and identify the variables dominated by these values it can be presented in terms of percentages.

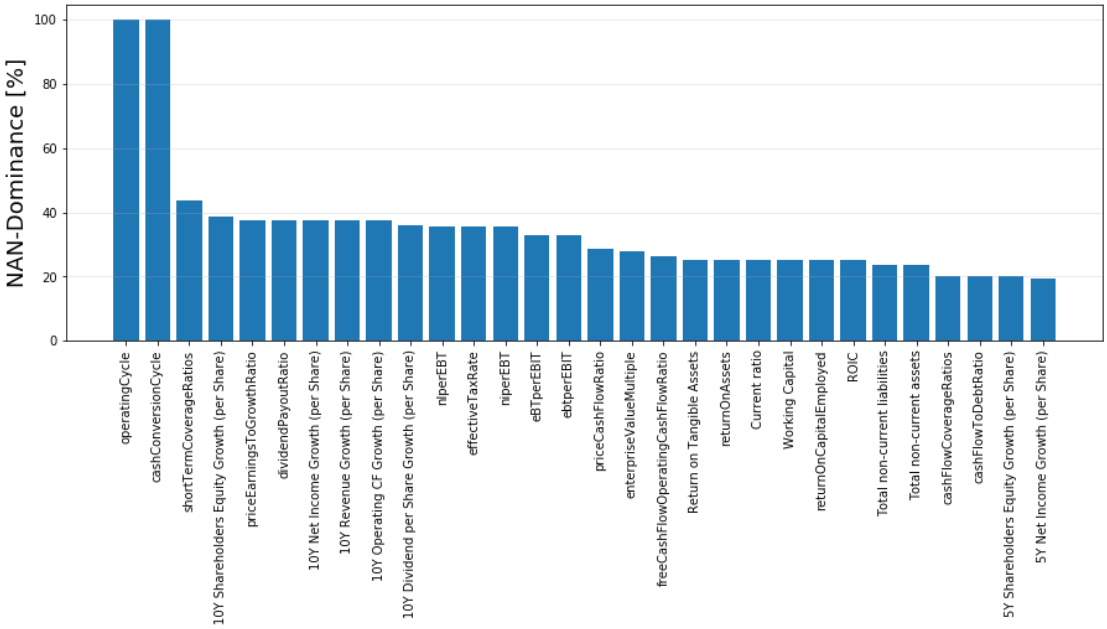


Figure 11. NAN in percentages.

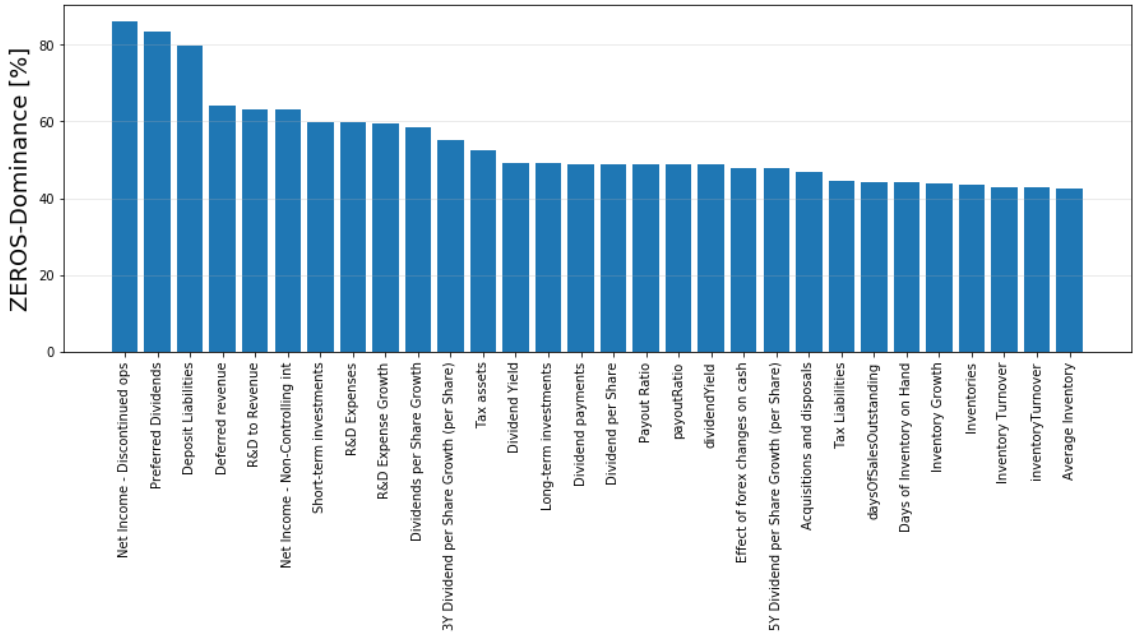


Figure 12. Zero values in percentages.

These figures shows that there is a need to adjust the quality of the data set, because some variables possess a high percentage of NAN and 0 values. It can be done by filling in the missing data, also called data imputing, or dropping indicators, which are too heavily dominated by these values. There are few methods to handle missing values: use mean from each variable, use the most frequent value from each variable, use interpolation, which employs a function that would match the data set, or even other methods like K-Nearest Neighbor, where missing values are set based on closes “k-points” or in other words neighborhood (VanderPlas, 2019). Based on how large this data set is and that it is separated by sectors, the reasonable choice would be to use the mean value. In order to better illustrate the effect of this, a correlation matrix plotting linear correlation between the variables can be of great use. Two figures below represents correlation between various financial indicators, including even those which are not being used in the empirical part. Figure 13 shows the situation before handling missing values.

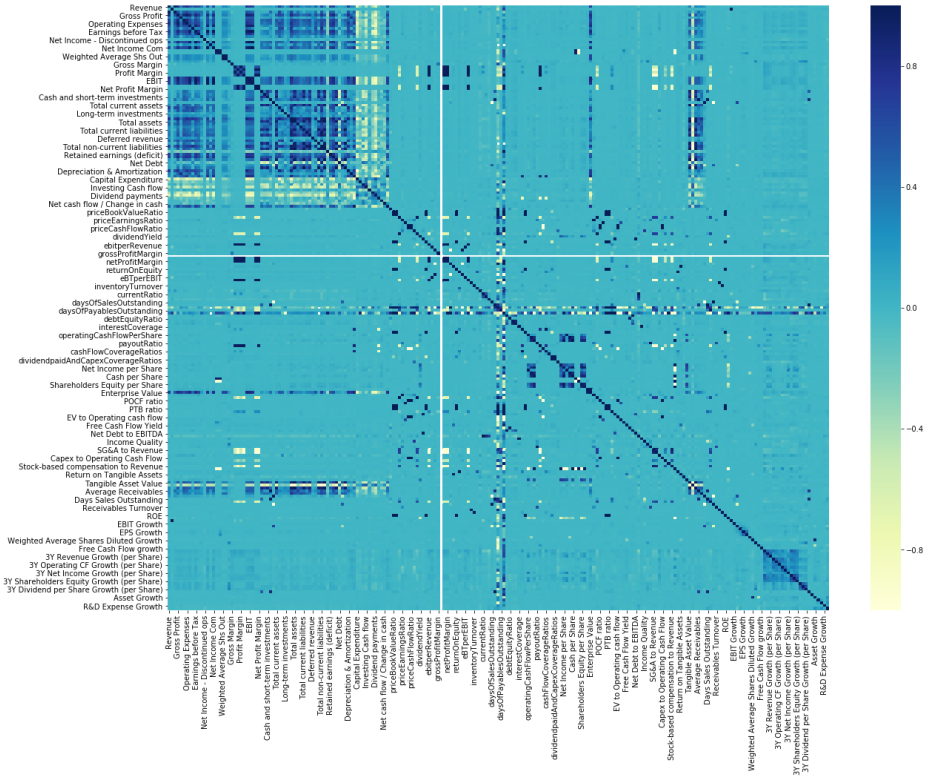


Figure 13. Data overview before cleaning.

Figure 14 is the correlation after handling missing values. It can be seen that there is a positive effect and correlation matrix looks more organic.

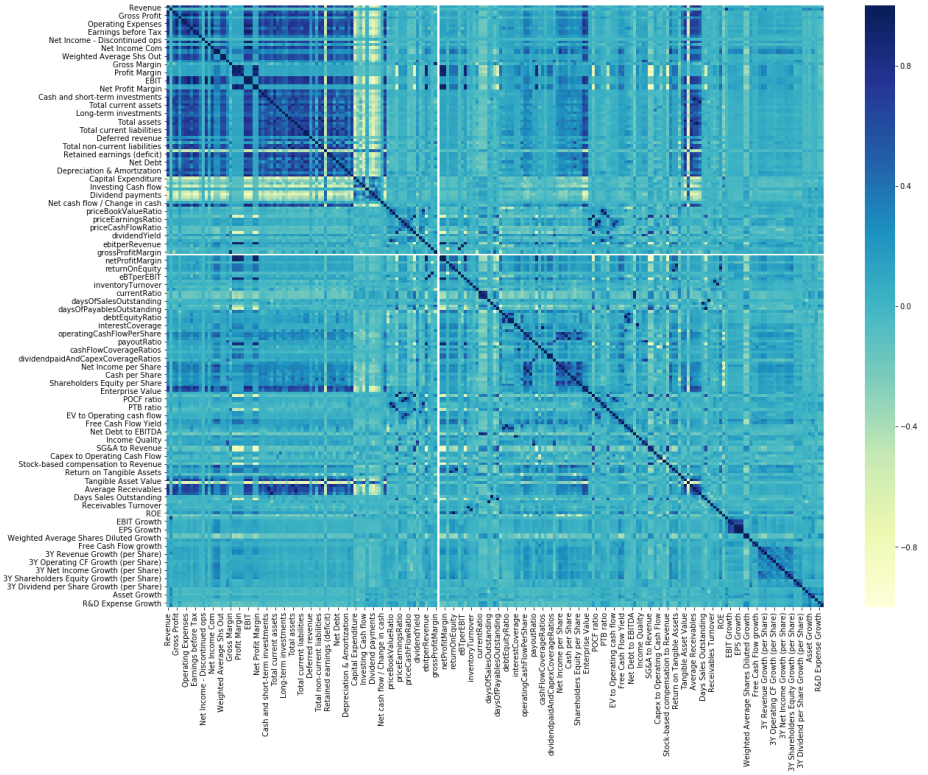


Figure 14. Data overview after cleaning.

Next step was feature engineering and selection.

Feature importance and engineering: machine learning models may appear as "black boxes," offering forecasts while providing no information on how they were created. Thus, there has been significant interest in developing mechanisms to either clarify the large trends modeled by these approaches or to justify specific predictions. Overall summaries of the influence of a specific input dimension on the models' forecasts can be provided by feature importance rankings. (Zhou, Hooker, 2020)

Hooker et al. (2019) also adds that in the context of machine learning, one of the most interesting and important problems is determining the effect of a particular input function on the prediction produced by a model. Recognizing which features are critical, aids in the improvement of our products, increases trust in model predictions, and segregates inappropriate variables (Hooker et al., 2019).

As per Pedregosa et al. (2011) the term "feature importance" refers to the process of assigning numerical values to input properties in predictive analytics. Feature value scores are critical components of predictive analytics since they can be used to optimize the model's output and provide context for the model and dataset. They may be used to define which features are most important for the study by using the relative values they are assigned. Under these methods, various forms of feature importance values are used. The most standard correlation ratings include Spearman's rankings or Pearson's correlation, for nonlinear and linear correlation, respectively. However, in this study a more advanced feature importance values are used, which can also be derived from linear models, or models based on decision tree, and permutation importance, as detailed in Pedregosa's et al (2011) work. Couple observations can be made regarding feature importance values:

Feature importance coefficients: following the fitting of a linear machine learning algorithm to the dataset, the coefficients of each input parameter can be obtained and expressed as feature importance values. The distinction is feasible due to the dataset's standardization and the variables' uniform size. This technique is used to extract feature importance values from Elastic-Net or Linear Regression models.

Feature importance with permutation method: this approach generates relative significance scores that are model-independent. After applying a model to the dataset, the forecast is performed multiple times to every feature in the dataset, yielding an average value for every input. This methodology is appropriate for models which don't have a native measure of function value (Altmann, 2010).

Kotsiantis (2007) also emphasizes that selecting features is basically the process of elimination and identification of as many redundant and irrelevant variables as possible. This

process helps to reduce the dimensionality of the sampled data and helps algorithms to perform calculations more efficiently.

Engineered features: for this study features were selected and engineered based on past scientific works and reasonability. The most popular, important, and some not-so-popular features were selected. Figure 15 represents the feature engineering part of the code.

```
def feature_engineering(x):
  x['Total Asset Turnover Ratio'] = x['Revenue'] / x['Total assets']
  x['Return on Total Assets Ratio'] = x['Net Income']/x['Total assets']
  features = ['Company Name', 'Current Ratio','Quick Ratio', 'Free Cash Flow Margin','Debt Ratio', 'Debt-To-Equity Ratio',
              'Cash Flow-To-Debt Ratio','Return on Tangible Assets','Total Asset Turnover Ratio',
              'Return on Total Assets Ratio','ROE','Gross Margin',
              'Net Profit Margin', 'EPS','EPS Diluted','PE ratio', 'PB ratio','Price Earnings-To-Growth Ratio', 'Sector','Class']
  x = x[features]
  return x
```

Figure 15. Feature engineering.

All separate data files of different years were merged into one data set. Figure 16 shows the description of the full data set between the years 2014-2018. Data set has 20 columns and 21040 rows where 2 columns are of object type, 1 integer and 17 float type.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 21040 entries, 0 to 4385
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Company Name                    21040 non-null  object
 1   Current Ratio                   21040 non-null  float64
 2   Quick Ratio                     21040 non-null  float64
 3   Free Cash Flow Margin           21040 non-null  float64
 4   Debt Ratio                      21040 non-null  float64
 5   Debt-To-Equity Ratio            21040 non-null  float64
 6   Cash Flow-To-Debt Ratio         21040 non-null  float64
 7   Return on Tangible Assets       21040 non-null  float64
 8   Total Asset Turnover Ratio      21040 non-null  float64
 9   Return on Total Assets Ratio    21040 non-null  float64
 10  ROE                             21040 non-null  float64
 11  Gross Margin                    21040 non-null  float64
 12  Net Profit Margin               21040 non-null  float64
 13  EPS                             21040 non-null  float64
 14  EPS Diluted                     21040 non-null  float64
 15  PE ratio                        21040 non-null  float64
 16  PB ratio                        21040 non-null  float64
 17  Price Earnings-To-Growth Ratio  21040 non-null  float64
 18  Sector                          21040 non-null  object
 19  Class                           21040 non-null  int64
dtypes: float64(17), int64(1), object(2)
memory usage: 3.4+ MB
```

Figure 16. Data set description.

Train and test split: as mentioned earlier it is critical to perform adequate sampling and data transformation to generate accurate and reliable forecasts. Following information refers to data sampling. To prevent data snooping prejudice, the data is divided into a test set and a training

set, as is usual in supervised learning. It is a common practice to use 80% of data for training part and 20% for testing. Stratified sampling was used in this study, in order to retain the same class ratios between the test and training sets. The stratified sampling technique splits the whole data into groups called strata. The appropriate volume of data is pooled to each group to guarantee that the research sample contains a comparable number of identical findings (Géron, 2017).

2.4. Prediction modelling

Regression and correlation: some of the simpler methods for analyzing relationship between two variables would be linear regression and correlation. Sir Francis Galton introduced the principle of linear regression in 1894. It is a mathematical test that is used to describe and measure the relationship between two variables in a data collection. Univariate regression methods like the Fisher's test, Chi-square, study of variance (ANOVA) and t-test do not enable for the use of additional covariates/confounders in studies. However, regression and partial correlation are the measures that enable the researchers to account for confounding variables when determining the relationship among the two variables (Chang, 2004).

If correlation measures the degree or intensity of a relationship among the two variables quantitatively, regression analysis mathematically defines this relationship. Regression analysis enables the prediction of the significance of a dependent variable using at least one independent variable's value. The correlation coefficient "r" is a non - dimensional integer with a value between -1 and +1. A value closer to -1 denotes an opposite or unfavorable relationship, while a value closer to +1 denotes a beneficial relationship. Pearson's correlation is used where the data has a regular distribution, while Spearman's correlation is used when the data has a non - normally distributed spread. The linear regression research employs the statistical equation $y = mx + c$ to define the line of optimal fit for relationship between y - dependent variable and x - independent variable. However, coefficient of regression, denoted by say r squared, indicates the degree of uncertainty in y as a function of x (Kumari and Yadav, 2018). Linear regression model was applied in this paper.

Classification with XGBoost classifier: supervised machine learning is a process of developing algorithms that are capable of producing general trends and predictions through predicting the outcome of future occurrences utilizing externally provided cases. Supervised learning classification algorithms are designed to classify data based on previous knowledge. Classification is a commonly used technique in data science challenges. Numerous promising methods, including instance-based techniques, logic-based techniques, rule-based techniques, and stochastic techniques, have been proposed to address these issues (Singh et al., 2016).

As per Tang et al. (2014) Classification is the process of grouping data. The main process of it is to determine to which of a collection of groups a new data point belongs to. This is done by using a training set of data comprising known-category observations. Numerous practical situations may be constructed as classification issues, for example classifying an email as “spam” or “non-spam”. This can be accomplished in two stages using the grouping technique. To begin, the method constructs a model, which is used for the class characteristic, as a feature of the dataset's other variables. Following that, it applies a beforehand constructed model to the newly created and previously unknown datasets in order to determine the relevant class of each record (Fernández-Gavilanes et al., 2016, Heydari et al., 2016).

According to Kotsiantis (2007) most important and distinguished machine learning techniques are: “Logic Based Algorithms: Decision Trees, Learning Set of rules; Perceptron-Based Techniques: Artificial Neural Networks; Statistical Learning Algorithms: Bayesian Networks, Instance-Based Learning; and Support Vector Machines”. Figure 17 represents the most common types of learning (Dey, 2016).

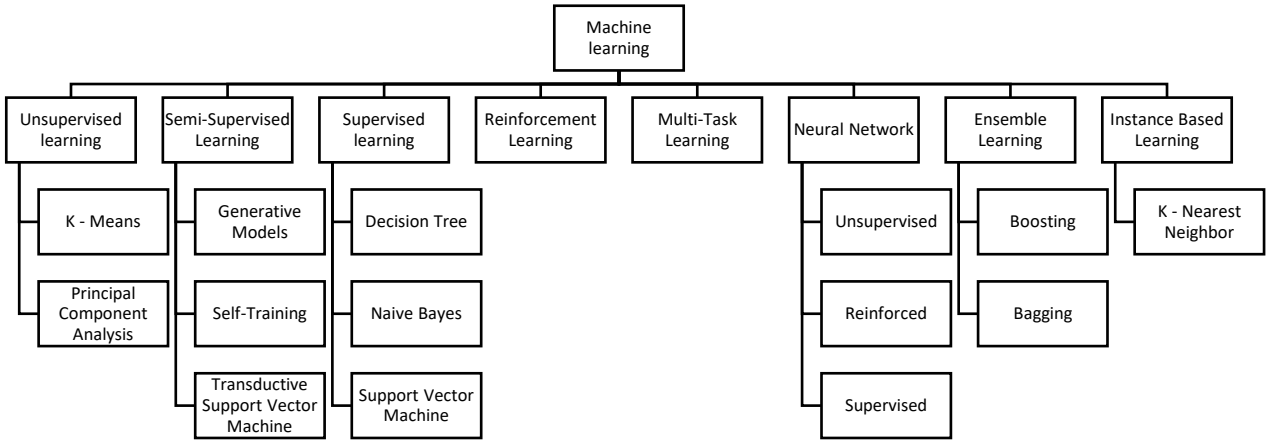


Figure 17. Machine learning techniques.

This study employed supervised machine learning technique for the classification problem, however XGBoost takes it to an ensemble boosting method. As decision-tree based models are mainly used for classification issues, a novel classifier from XGBoost library was chosen. “XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework.” (XGBoost Documentation, 2020).

The Decision Tree algorithm is a member of the supervised learning algorithm category. In addition to tackling regression problems, the decision tree algorithms, can also be used to solve classification problems. The aim of a Decision Tree is to develop a training model capable of predicting the class or importance of a target variable by inferring basic decision rules from

previous information (training data). In Decision Trees, the first step is to predict a class mark for an observation at the tree's base. The attributes of the root then are compared to the observation. Based on the comparison, the model follows the tree branch which corresponds to that value and then goes onto the next node (Alsagheer et al., 2017).

Out of the machine learning techniques in use today, gradient tree boosting is one that excels in a variety of applications. Tree boosting has been demonstrated to achieve excellent performance on a variety of widely used classification benchmarks (Chen and Guestrin, 2016).

Boosting is an ensemble learning method that is used to reduce variance and bias. Boosting takes a group of learners which are weak and transforms them into a single powerful learner. A learner categorized as weak is a classification algorithm that has a very low correlation with classification that is defined as true. A strong learner, on the other hand, is a form of classifier that is highly associated with true classification (Zhou). In other words, it is a method in which new models are added to compensate and fix for previous models' errors. These models are added in a sequence till no more enhancements are possible. Gradient boosting is a method for predicting those additional errors of previous models, then combining them to make the final forecast. The term "gradient boosting" refers to the fact that it employs a gradient descent algorithm to avoid the errors when adding new models (Chen and Guestrin, 2016). XGBoost is known for its excellent execution speed and model performance.

Boosting ensemble technique can be simplified and summarized in three steps models (Chen and Guestrin, 2016):

1. to predict the target variable y, the first model F0 is created
2. a new model h1 is attached to the residuals from the prior step
3. F0 and h1 are merged to produce F1, the boosted variant of F0. The mean squared error from F1 is then lower compared to F0

   In mathematical formula it can be expressed as follows:

   $F1(x) = F0(x) + h1(x)$

In order to increase the performance of F1, it models looking at the residuals of F1, and creates a model F2:

$F2(x) = F1(x) + h2(x)$

This technique can be repeated for "n" iterations, until the residual losses are reduced as much as possible:

$Fm(x) = Fm\text{-}1(x) + hm(x)$

Going into more specifics regarding the XGBoost model and machine learning in general, here is a general principle or objective function of supervised learning:

$$\text{obj}(\theta) = L(\theta) + \Omega(\theta)$$

Where $L(\theta)$ is training loss and is $\Omega(\theta)$ regularization. The aim of the function is to find the best parameters θ. Training loss shows how accurate and how predictive the model is in regards to the training set. While regularization is the complexity of the model. It should help to avoid overfitting, which in other words is training data "too well", where random variances or noise is picked up by the model as training.

Now after the general principle, it is important to introduce Decision Tree Ensembles function. The optimized objective function can be described as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

where K is the amount of trees, y is the target variable, f is a function inside the functional space, $\mathcal{F}$ would be the collection of all possible CARTs (classification and regression trees), and x are the features.

Next step is boosting of the trees, which can be expressed as:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

Where the structure of the formulas can be interpreted as additive training. Each tree is learned in a different step. First, model fixes prior mistakes, then adds the new tree. Here f_i represents what is needed to learn and prediction value is expressed at step t as $\hat{y}_i^{(t)}$.

As training part is covered, second part is regularization or complexity. In XGBoost it is expressed as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Where *w* represents the vector of values on leaves, *j* assigns each data point to correct leaf, and T is the number of leaves.

After re-formulating the whole tree model, the objective value can be expressed as:

$$\text{obj}^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$

These expressions are defining the measurement of how good the tree is. Model also needs to learn the structure of the tree, which can also be expressed mathematically as:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

The last formula can be analyzed as follows: first it calculates the value of new left leaf, then the value on the new right leaf, original leaf, and regularization on the new leaf. This formula also means that if the gain is lower than γ, then the best choice is to not add that branch to the model. Taking into consideration all described formulas, Gradient tree boosting algorithm can be expressed as following:

Algorithm 4: Gradient tree boosting

Input : Data set $\mathcal{D}$.
A loss function L.
The number of iterations M.
The learning rate η.
The number of terminal nodes T

Initialize $\hat{f}^{(0)}(x) = \hat{f}_0(x) = \hat{\theta}_0 = \arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \theta)$;

for $m = 1,2,..,M$ **do**

  $\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}^{(m-1)}(x)}$;

  Determine the structure $\{\hat{R}_{jm}\}_{j=1}^{T}$ by selecting splits which maximize $Gain = \frac{1}{2}\left[\frac{G_L^2}{n_L} + \frac{G_R^2}{n_R} - \frac{G_{jm}^2}{n_{jm}}\right]$;

  Determine the leaf weights $\{\hat{w}_{jm}\}_{j=1}^{T}$ for the learnt structure by $\hat{w}_{jm} = \arg\min_{w_j} \sum_{i \in \hat{I}_{jm}} L(y_i, \hat{f}^{(m-1)}(x_i) + w_j)$;

  $\hat{f}_m(x) = \eta \sum_{j=1}^{T} \hat{w}_{jm} \mathrm{I}(x_i \in \hat{R}_{jm})$;

  $\hat{f}^{(m)}(x) = \hat{f}^{(m-1)}(x) + \hat{f}_m(x)$;

end

Output: $\hat{f}(x) \equiv \hat{f}^{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$

Figure 18. Gradient tree boosting algorithm.

The aim of XGBoost is to push the machine learning computation capabilities to the extreme in terms of speed and efficiency (Chen and Guestrin, 2016).

ROC score and ROC curve: in order to help visualize the results, the receiver operating characteristic curve (ROC) was built. ROC is used to investigate and quantify the interaction between a binary classifier's sensitivity and accuracy. Sensitivity, or true positive score, quantifies the proportion of correctly classified positives; accuracy, or true negative rate, quantifies the proportion of correctly classified negatives. The roots of ROC research can be traced back to signal detection theory. A detection problem, in simplified sense, helps determine the significance of a binary signal polluted with random noise. Without any additional data, the logical judgment threshold will be in the middle between these two signal values. In case the noise spread is zero-

centered and symmetric, the sensitivity and accuracy at this threshold are equal, which implies that the related operational point on the ROC curve is situated at the junction with the falling diagonal (Flach, 2016).

SHAP value analysis: moreover, for visualization of the results and in order to explain the "black box" of machine learning results, Shapley Additive Explanation (SHAP) tool was used as well. In case of Casalicchio's et al., (2019) study, SHAP allowed the authors to construct a very complex XGBoost machine learning model, which was capable of generating extremely precise forecasts and allowed for individual-level understanding of the variables that contributed to the model results. SHAP is a machine learning model explanation technique, which helps to explain individual forecasts. It is based on the game theoretically optimal Shapley Values. Which states that a prediction can be understood in the following way: each feature is interpreted as a "player" in game in which the prediction represents the prize or payoff. Shapley Values is basically a technique from a game theory – which indicates how to distribute the prize between the features in a rational manner.

3. RELATIONSHIP BETWEEN STOCK PERFORMANCE AND FINANCIAL RATIOS USING MACHINE LEARNING EMPIRICAL RESULTS

3.1. Exploratory analysis

The empirical findings chapter is split into three subsections: overall analysis findings, linear regression and correlation findings, and classification modeling findings using the XGBoost classifier.

The exploratory stage of this data analysis is performed with an aim to illustrate the dataset's overall state and present its story. This chapter highlights the key results from the price variance distribution study and illustrates the link between financial indicators and stock price variation using visualizations to aid in prediction modeling which is discussed in further chapters.

Distribution of stock price variance

The depth of the previous recession in the United States in 2008–2009 prompted both the central bank and government to pursue expansionary monetary and fiscal policies. In terms of monetary policy, the US Federal Reserve, in addition to its ultra-low federal funds rate, has announced a quarterly USD 85 billion bond-buying program, which was targeted at limiting long-term interest rates. Few previous years' stock prices represent predicted cash flows (profits) discounted by relevant interest rates. Discounted cash flows were large, mainly because of very low interest rates, which justifies the recent stock price gains. Since March 2009, the US stock markets have been trending higher at quite high rates of return (Huang et al., 2016). If we look at the recent distribution of stock price variance, the trend is indeed positive. Figure 1 shows the distribution of stock price variance for year 2015. Figure 19 data represents variances between -100 and 500 percent.

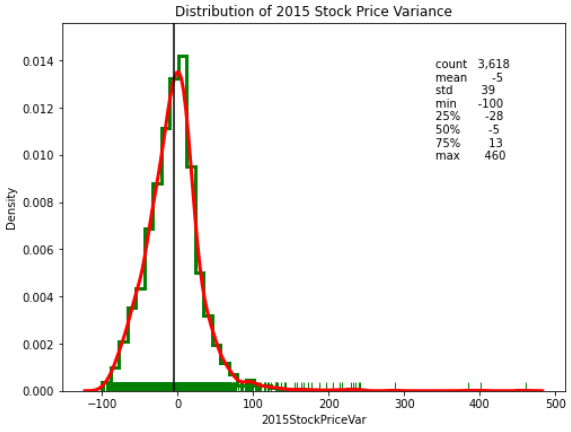


Figure 19. Distribution of 2015 Stock Price Variance.

From figure 19 it can be interpreted that year 2015 was rather calm for US stock market, as stock price variance distribution is quite normal. There is no clear or dominant negative or positive density. Mean was -5%, which indicates slightly negative climate overall for stock prices that year. Looking at the quartiles, same story is evident, with lower quartile being at -28%, middle at -5% and upper at 13%. Dispersion or standard deviation, which is the spread of observations around the central tendency was at 39%, considering whole sample between -100% and 500% stock price variances. However, the quality of information absorbed can be increased if we zoom in a little bit and look only at the variances lower than 100%, which is still a lot of room even for unreasonable growth, but it provides a more reasonable picture of overall situation.

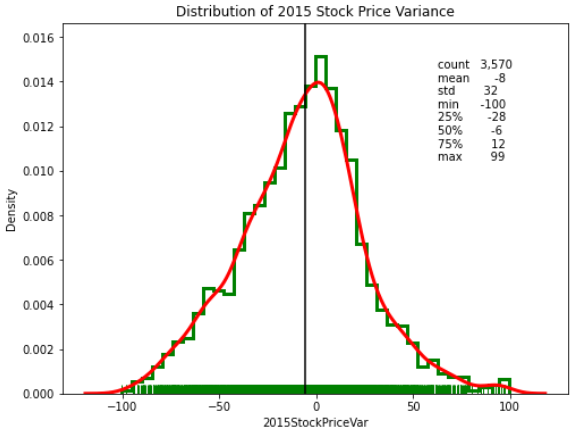


Figure 20. Distribution of 2015 Stock Price Variance, no outliers.

Figure 20 represents stock price variances lower than 100% of previously described figure. As we can see the number of observations decreased only in a very insignificant amount, which is a mere 1%. It can be seen that there was a peak of density around 0% or even slightly more than 0%, however the steps were higher at negative side and the tail of distribution falls faster at the negative side. Confirming what was shown in previous figure, that year 2015 was slightly trending negative for US stocks. Mean was -8%, lower quartile -28%, middle quartile -6%, and upper quartile 12%. Standard deviation stood at 32%. Figure 21 presents information for year 2016 in the same manner.

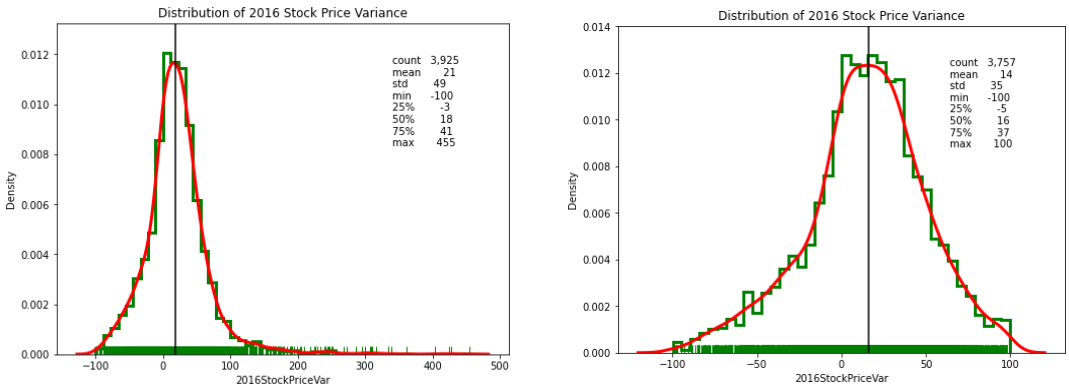


Figure 21. Distributions of 2016 Stock Price Variances.

From figure 21 it is evident that stock price variances were trending to positive side at the year 2016. Mean was 21%, lower quartile -3%, middle quartile 18%, and upper quartile 41%. However, if we look into the zoomed in picture, we can see slightly more characteristics of that year. Mean was 14%, standard deviation 35%, lower quartile -5%, middle quartile 16%, and upper quartile 37%, which indicates a positive year for stock price variances overall. Same hypothesis can be supported by looking at the density – it is higher on the positive side, meaning more stocks experienced positive gains during year 2016. Figures 22 shows data for year 2017.

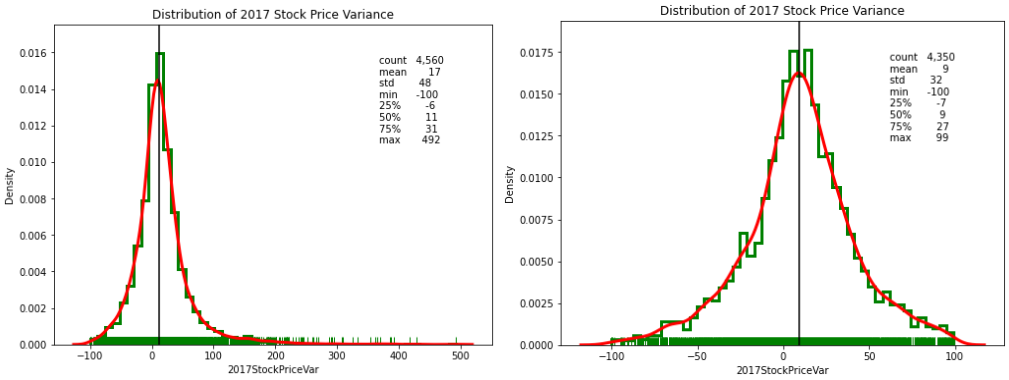


Figure 22. Distributions of 2017 Stock Price Variances.

Year 2017 was highly positive for US stock market. Mean was at 17%, which is quite high. Quartiles can only confirm that, lower being at -6%, middle at 11% and upper at 31%. Standard deviation was 48%. If we look only at the stocks, which experienced gains between -100% and 100% the picture is even clearer. Mean was 9%, lower quartile -7%, middle quartile 9%, upper quartile 27%. Density does not really give a clear indication, which side was more dominant.

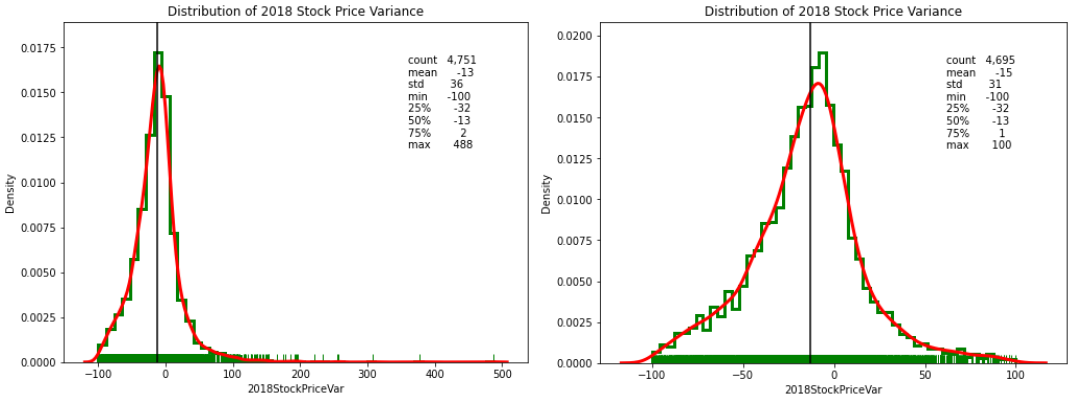


Figure 23. Distributions of 2018 Stock Price Variances.

Figure 23 shows the data in the same manner for year 2018. We can see that the mean was negative with -13% that year, which indicates a downward trend for stocks. Looking at the

quartiles, same conclusion can be made. Lower quartile stood at -32%, middle at 13% and upper only at 2%. Standard deviation was 36%. Regarding density, the tail of course was larger on the positive side as stocks can go infinite amount to the positive side and only -100% to the negative. However, looking at the steps of the density it can be seen that the negative side is slightly denser, meaning that probably there were a lot more negative price variances that particular year. From looking at the zoomed in figure we can only confirm that. Mean was -15%, standard deviation 31%, lower quartile -32%, middle quartile -13% and upper only 1%. Stock price variances were heavily based on the negative side. Even though it can be observed that there is a density peak around 0%, the steps are higher at the negative side of the distribution.

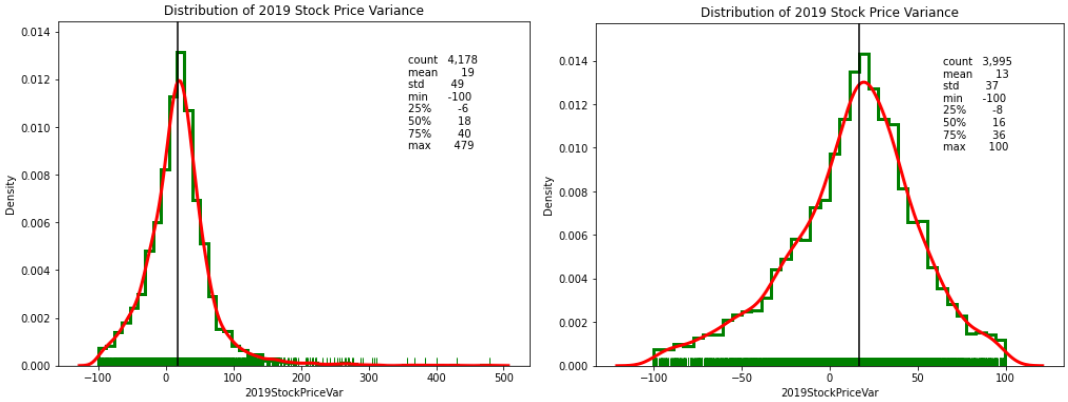


Figure 24. Distribution of 2019 Stock Price Variances.

Figure 24 presents stock variance data for year 2019. Even in such large samples the central tendency was positive, mean was around 19%. While the dispersion can be interpreted as quite large, which was 49%. Looking at the quartiles, it is also evident that the trend was positive. Lower quartile was only -6%, middle 18% and upper 40%. Also, it can be seen that the density was trending towards to positive side of the mean. If we look at the data in the figure without higher gains than 100%, it is evident that the mean was still positive with 13%, which is relatively good for stock market. Standard deviation decreased, because we have less dispersion now in the sample, and was at 37%. However, as expected lower quartile decreased a little bit, it went from -6% to -8%, however middle and upper quartiles are still highly positive, being respectively at 16% and 36%. If we look at the density, it can be seen that there is a "tail" at the negative side, which means that it is not so dense there, in other words – in comparison to all, not a lot of companies experienced negative price variance. It can be said that the year 2019 was a relatively positive year for US stock market as a whole.

Stock price variance based on sectors

Price variance was also analyzed in terms of different sectors in US stock market. Following figure represents that information for year 2015 and all the rest of the years can be

found in appendix. For reasonability purposes data in these figures only represents variances from -100% to 100%. Outliers with gains of more than 100% skewed the data to one side and did not provide any valuable input into analysis, particularly in this section.

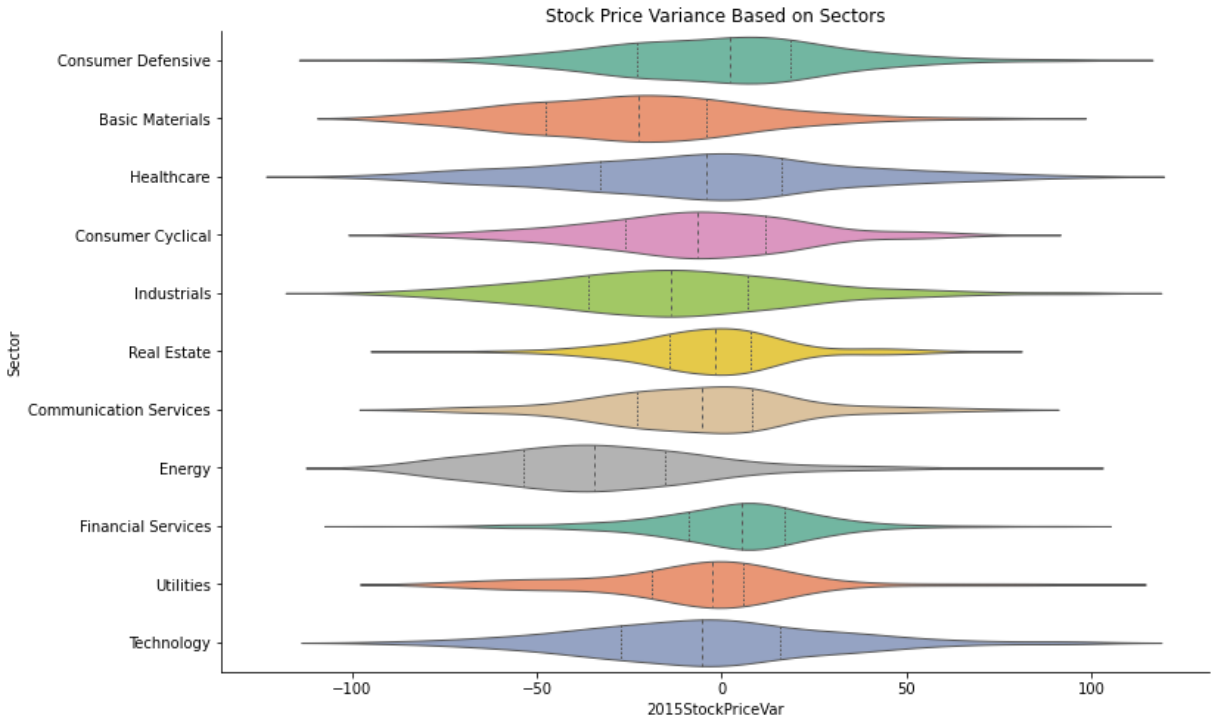


Figure 25. Stock price variance based on sectors.

It can be observed that in year 2015 most of the Energy sector specifically experienced rather negative stock price variances. Same can be stated about Basic Materials sector, its distribution was also heavily based on the negative side. However, Financial Services and Consumer Defensive that year overall had more positive distribution of stock price variance. Similar trends can be observed every time there is a bigger dip in the overall market due to some more or less global trigger. Defensive stocks are called defensive for a reason, investors tend to flock to them when there is a turmoil in the market. Looking at the year 2016 distribution it is evident that almost all sectors had their stock price variances based in positive percentages. Except for one – Healthcare, its distribution was actually relatively widely spread, but the mean is in the negative percentages. During 2017 only Energy sector had its mean in the negative percentages, and whole distribution shifting in the negatives as well. Most of other sectors experienced positive variances overall. Financial Services sector's distribution looked more tightly squeezed if compared to previous years, indicating slight consolidation in stock returns between financial service companies in US. 2018 was overall a more negative year for stock market and it is evident from the distributions, only Utilities and Consumer Defensive had their mean around 0 or slightly

positive, which again supports defensive stocks theory. Finally, looking at the last year can see that 2019 was again overall a more positive year for stock prices in US. Distributions were leaning to positive percentage area and most of the sector's means were in the positives as well. In particular Real Estate and Utilities experienced a more positive year compared to other sectors.

Visualized relationships

Continuing explanatory analysis, visualizations of the relationship between financial indicators and stock price variance were made. Ratios were grouped into 5 categories: liquidity – figure 26, leverage – fig. 27, operational – fig. 28, profitability – fig. 29 and valuation – fig 30. Y axis shows stock price variance and X axis shows the corresponding financial ratio. For research and reasonability purposes only year 2019 will be presented in this paper.

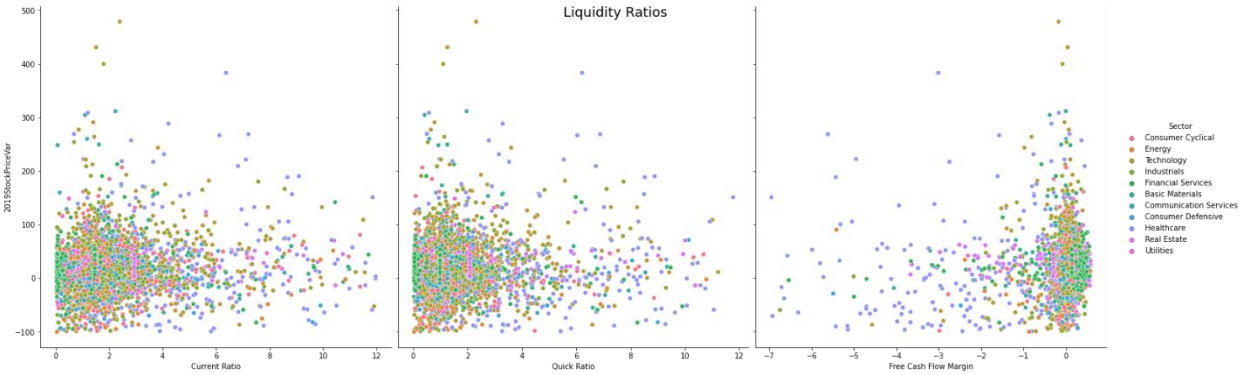


Figure 26. Liquidity ratios.

From liquidity ratios it can be observed that Healthcare and Technology have higher spread of distribution in Current Ratio compared to other sectors. Three observations can be seen from Technology sector between 400-500% variances and their Current Ratio was around 2. Similar situation is evident with Quick Ratio, Healthcare and Technology sectors spreading wider in terms of stock prices and financial ratio. Majority of observations are between -100% and 100% variances, and 0-4 in terms of Quick Ratio value. Looking at the Free Cash Flow Margin it can be said that the majority of observations are between -1 and +1 in terms of financial ratio value. Healthcare here also dominates in the spread of financial ratio value. An explanation to that could be that many pharmaceutical companies bet on one drug success and until then they are not generating any profits. Most of the new pharmaceutical companies lose money for many years even if they have some revenues, because of very high R&D costs and long drug development times.

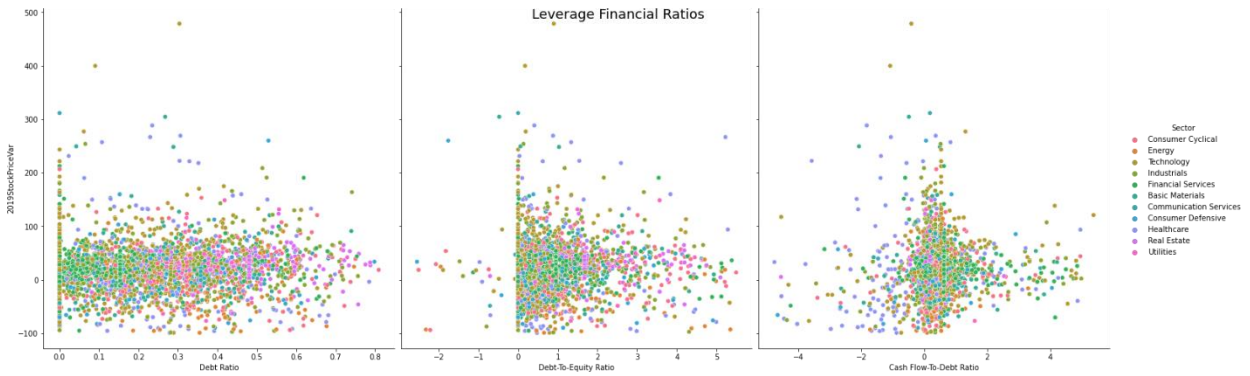


Figure 27. Leverage ratios.

Looking at the leverage financial ratios, it can be seen that some of the companies are near or at 0 in terms of Debt Ratio. This can be explained by the fact that some companies have periods when they are free of debt. Distribution of Debt Ratio in terms of stock price variance looks rather normal, Healthcare and Technology again dominating the top percentages of price variances. It can also be emphasized that Real Estate is rather distinctly spreading towards the upper quartiles of Debt Ratio. Debt-To-Equity ratio shows slightly different picture, few companies experienced negative ratios. Negative Debt-To-Equity ratios indicates that the company can be extremally risky and nearing bankruptcy. Such companies were from various sectors as can be seen, however they were just a few. Higher Debt-To-Equity ratio values can also be observed in few sectors, but particularly again in Real Estate. Finally, Cash Flow-To-Debt ratio provides rather normal distribution, with Healthcare and Technology dominating negative values and Financial Services positive ones.

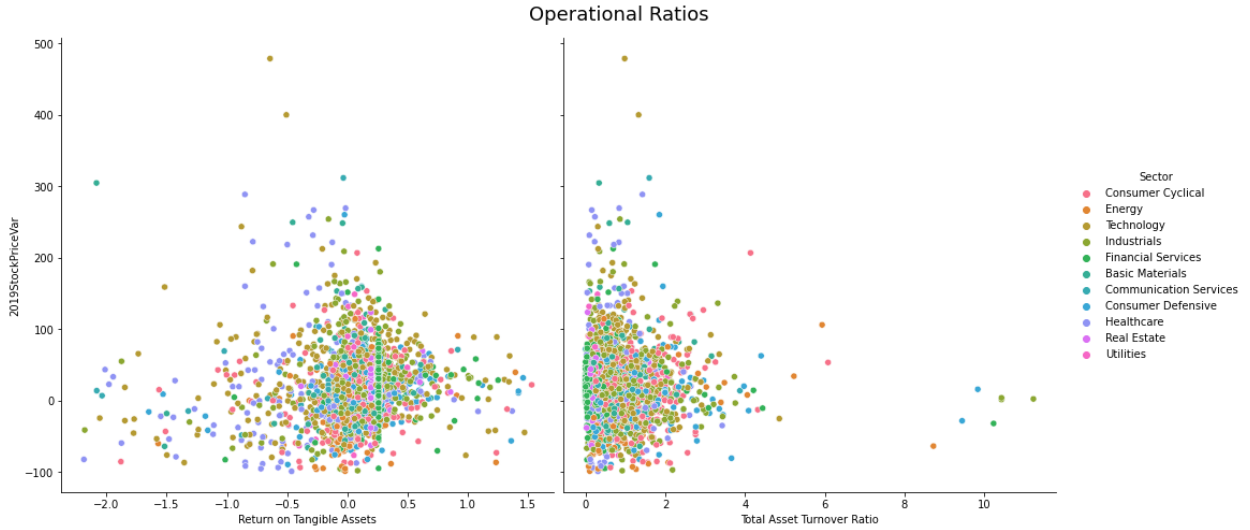


Figure 28. Operational ratios.

Couple operation ratios were analyzed. In terms of Return on Tangible Assets ratio Healthcare once again had the highest spread. Most of the distribution stacked between -0.5 and 0.5 on Return on Tangible Assets and around -50% and 100% in regards of stock price variances.

Looking at the Total Asset Turnover Ratio it can be seen that there were few outliers with very high ratio valuations, however the returns of stock price for those companies were not significant. Most of the distribution looks to be between 0 and 2 in terms of ratio valuation.

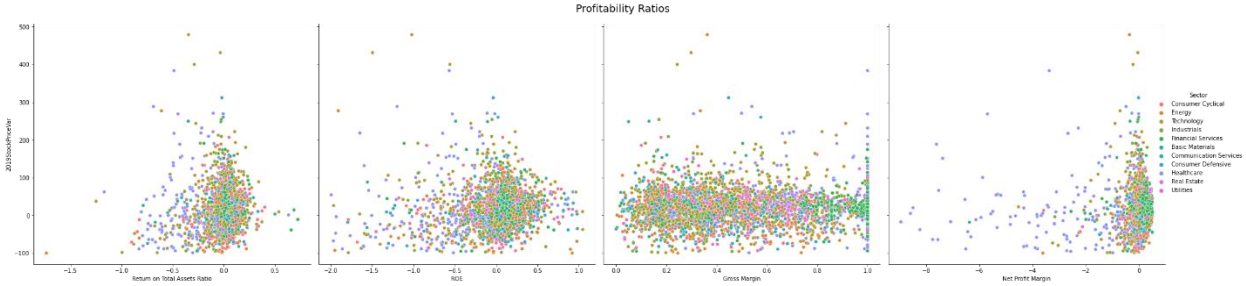


Figure 29. Profitability ratios.

Regarding profitability ratios, Healthcare dominates the negatives values for Return on Total Assets ratio. While most of observations were between -0.5 and 0.5. Looking at Return on Equity (ROE) one can make an observation that the stock price variance and ROE has quite the normal distribution and relationship does not have any significant characteristics. However, regarding Gross Margin, many Financial Services companies experienced around 100% of that ratio valuation. Investment management companies, especially new ones, tend to have good returns and good markups on their services. Net Profit Margin looks distributed around a -1 and 1. Only Healthcare sector is evidently spread in terms of that ratio, because of their business model. However, as can be seen from figure 29 the stock price variance tends to be higher the more positive one's Net Profit Margin is.

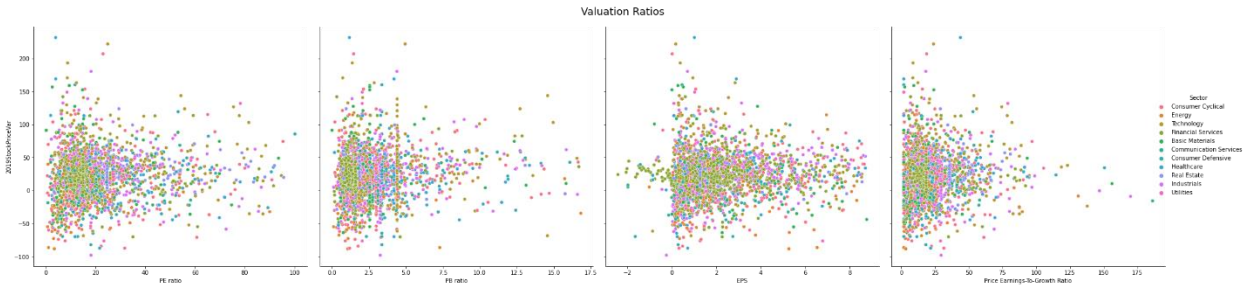


Figure 30. Valuation ratios.

Valuation ratios in Figure 30 include PE ratio, PB ratio, earnings per share (EPS) and Price Earnings-To-Growth Ratio. For PE ratio distribution of observations are mainly based between 0 and 40 and -50% to 50% in terms of stock price variance. PB ratio very similarly to PE ratio has a normally distributed relationship with stock price variance. Few companies experienced negative EPS, indicating loss of money. Most of those companies were from Technology sector, as can be

explained by rising number of Technology start-ups. Price Earnings-To-Growth (PEG) ratio also shows rather normal distribution with few outliers being from various sector.

In summary, all five years' price variances were typically distributed with a positive or negative bias depending on the year's performance. No distributional anomalies were identified. Sector-specific stock price variance investigation revealed that sectors do react and move slightly differently. Additionally, there are so-called "defensive" sectors in which stock values do not plunge when all other sectors do. However, based on the typical stock price distribution, the "conservative" sectors are unlikely to generate significant profits. Then, when the relations were illustrated, it became clear that the healthcare and technology industries dominate in terms of growth and risk. The explanation for this might be considered to be the increased number of new businesses or start-ups and their potential in these two industries.

3.2. Linear regression and correlation examination

Correlation and linear regression are two strategies for determining the connection between two parameters. Correlation describes the strength of a linear connection between two variables quantitatively, while regression describes the connection as an equation (Kumari and Yadav, 2018). Regression is used to forecast a quantitative value and to conduct actions on a dataset in which the goal values have already been established. Additionally, the result may be expanded by adding other data. The relationships established by regression between prediction and endpoint values might form a pattern. This pattern may be used to other datasets whose goal values are unknown. (Gharehchopogh et al., 2013) As a result, the data required for regression are divided into two sections: one for creating the model and another for testing the model. In this part, linear regression was used to do the study. To begin, data was separated into two categories: training and testing. Following that, the training portion was utilized to begin the analysis and define the model.

For linear regression modeling Simple Imputer and Standard Scaler were used from Sklearn machine learning library. Figure 31 clarifies that the model which was fit was Linear Regression.

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
```

```
LinearRegression()
```

Figure 31. Model fit - linear regression.

Different data and different statistical models not necessarily can work well with each other and blindly applying them to the study is not reasonable. Therefore, before analyzing any results, it is important to check how well the models fits the data, how well it explains the data and is there any errors. Couple methods how to do that is errors statistics and explainability check. Error statistics were checked first using Mean Absolute Error (MAE) and Mean Square Error (MSE) methods. The absolute error is the difference seen between predicted and actual values. MAE shows us how much inaccuracy we may anticipate on average from the projection. In this case, "error" refers to measurement ambiguity or the discrepancy between the estimated and correct or true values. Unlike other error statistic methods, instead of just average, MAE calculates average errors using the absolute values of faults, which results in more logical average errors. While, MSE is regarded as most often used loss function for regression. It is the product of the squares of the distances between model predictions and target or real variables. MAE is useful if data set has many outliers, while MSE should give a more stable calculation. (Davydenko, 2014). The lower the value the better, however there is no universal threshold, which could act as a benchmark for good model, it is always advised to have the data set in mind when judging error sizes.

Figure 32 shows that Linear Regression model used in this study produces 1539.23 MSE for train set and 1272.22 MSE for test set. Also, 24.74 MAE and 23.74 MAE for train and test sets respectively. These values are not low even for this dataset and these error statistics can already act as a red flag, that something is not good with the model and this data set.

```
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
mse_train = mean_squared_error(y_train, y_train_pred)
mse_test = mean_squared_error(y_test, y_test_pred)
mae_train = mean_absolute_error(y_train, y_train_pred)
mae_test = mean_absolute_error(y_test, y_test_pred)
print(f'MSE for train set is: {mse_train:,.2f} ')
print(f'MSE for test set is: {mse_test:,.2f} ')
print(f'MAE for train set is: {mae_train:,.2f} ')
print(f'MAE for test set is: {mae_test:,.2f} ')
```

```
MSE for train set is: 1,539.23
MSE for test set is: 1,272.22
MAE for train set is: 24.74
MAE for test set is: 23.74
```

Figure 32. Linear regression's error statistics.

Next step was check of model explainability. R squared is a measure to do that. It is a statistic that indicates the amount of variation that can be explained by an independent variable in a regression model. Figure 33 shows results of the check.

```
from sklearn.metrics import r2_score
y_train_class_pred = model.predict(X_train_class_sc)
y_test_class_pred = model.predict(X_test_class_sc)
r2_train_class = r2_score(y_train, y_train_class_pred)
r2_train_class
```

```
0.035507288207150656
```

Figure 33. Linear regression's R squared.

Returned R square score was a 3.55%, which is extremally low. It means that only 3.55% of data sets stock price variance could be explained using this Linear Regression model. No further steps were taken regarding Linear Regression, because any results generated by this model could not be trusted or used for any conclusion regarding the relationship between financial ratios and stock price variance.

Further, correlation between the variables was investigated. Correlation between the variables is shown as a heatmap also using different sized squares depending on strength of relationship. Color also denotes relationship, blue being the strongest with a score 1 and red the weakest with -1. Due to better visualization purposes the percentage values are not displayed in the below Figures, however they are described.

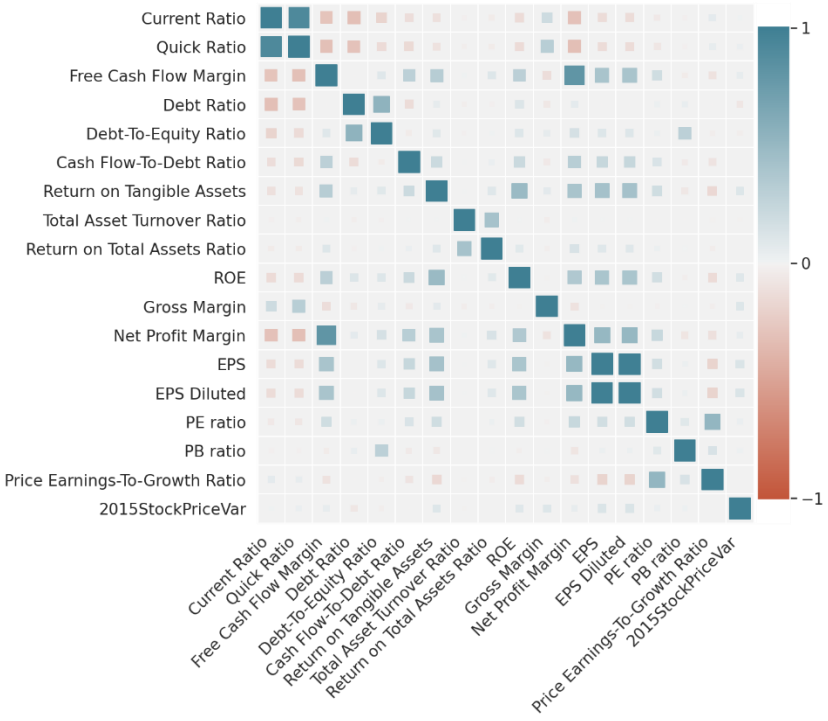


Figure 34. Correlation heatmap analysis, year 2014.

Figure 34 shows the correlation between variables in year 2014 and stock price variance in year 2015. All 17 variables had quite weak correlation with stock price variance. Following results were observed: Current Ratio 0.013, Quick Ratio 0.036, Free Cash Flow Margin 0.052, Debt Ratio -0.067, Debt-To-Equity Ratio -0.025, Cash Flow-To-Debt Ratio 0.0039, Return on Tangible Assets 0.022, Total Asset Turnover Ratio -0.0052, Return on Total Assets Ratio 0.0017, ROE 0.1, Gross Margin 0.11, Net Profit Margin 0.052, EPS 0.12, EPS Diluted 0.12, PE Ratio 0.043, PB Ratio 0.025. Price Earnings-To-Growth Ratio 0.0013.

However, there were some stronger correlations between other variables. The strongest few positive relationships in terms of correlation were recorded between: EPS and EPS Diluted 1.00, Quick Ratio and Current Ratio 0.92, Net Profit Margin and Free Cash Flow Margin 0.82, Debt Ratio and Debt-To-Equity Ratio 0.55, Price Earnings-To-Growth Ratio and PE Ratio 0.52, EPS and Net Profit Margin 0.51. All the other variables had correlation values lower than 0.50. These findings are generally consistent with those of the previous scientific literature.

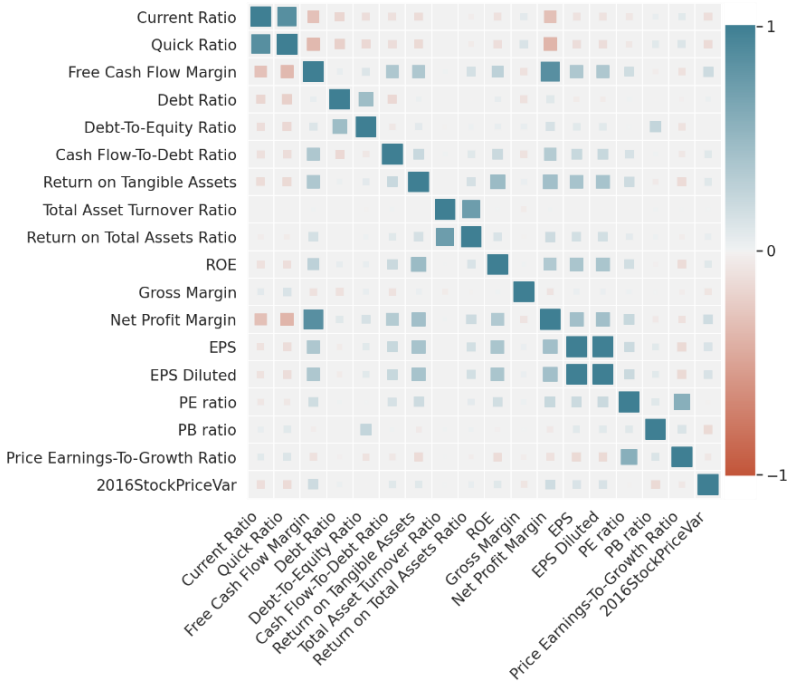


Figure 35. Correlation heatmap analysis, year 2015.

Figure 35 shows the correlation between variables in year 2015 and stock price variance in year 2016. All 17 variables had quite weak correlation with stock price variance. Following results were observed: Current Ratio -0.12, Quick Ratio -0.14, Free Cash Flow Margin 0.21, Debt Ratio 0.031, Debt-To-Equity Ratio 0.0022, Cash Flow-To-Debt Ratio 0.095, Return on Tangible Assets 0.0022, Total Asset Turnover Ratio 0.011, Return on Total Assets Ratio 0.048, ROE 0.094, Gross Margin -0.076, Net Profit Margin 0.2, EPS 0.14, EPS Diluted 0.14, PE Ratio -0.0056, PB Ratio -0.15. Price Earnings-To-Growth Ratio -0.067. The strongest few positive relationships in

terms of correlation were recorded between: EPS and EPS Diluted 1.00, Quick Ratio and Current Ratio 0.87, Net Profit Margin and Free Cash Flow Margin 0.86, Return on Total Assets Ratio and Total Asset Turnover Ratio 0.74, Price Earnings-To-Growth Ratio and PE Ratio 0.59. All other variables had correlation values lower than 0.50.

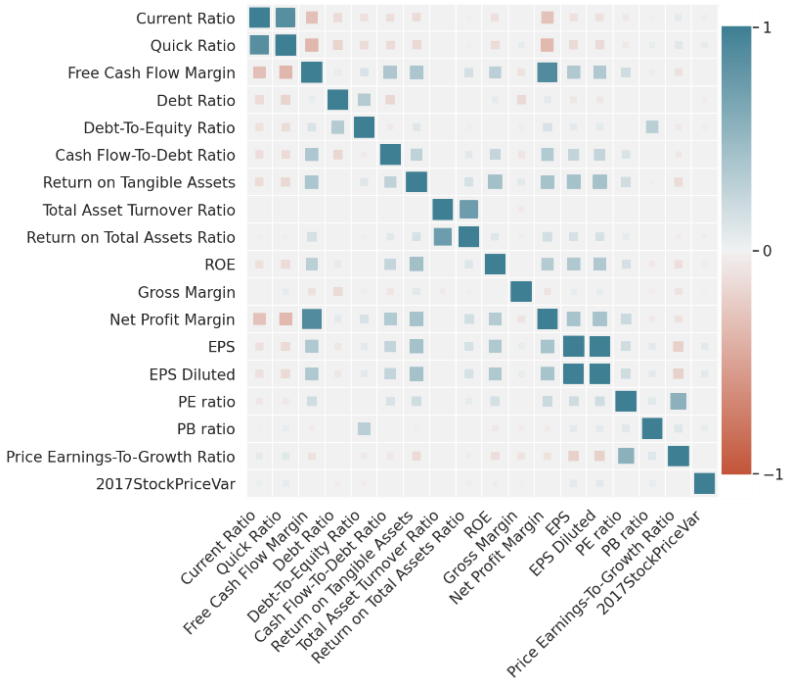


Figure 36. Correlation heatmap analysis, year 2016.

Figure 36 shows the correlation between variables in year 2016 and stock price variance in year 2017. All 17 variables had quite weak correlation with stock price variance. Following results were observed: Current Ratio 0.042, Quick Ratio 0.059, Free Cash Flow Margin -0.0082, Debt Ratio -0.036, Debt-To-Equity Ratio -0.029, Cash Flow-To-Debt Ratio 0.0073, Return on Tangible Assets 0.042, Total Asset Turnover Ratio 0.019, Return on Total Assets Ratio 0.037, ROE 0.039, Gross Margin 0.021, Net Profit Margin 0.0073, EPS 0.079, EPS Diluted 0.079, PE Ratio -0.011, PB Ratio 0.058. Price Earnings-To-Growth Ratio -0.0044. The strongest few positive relationships in terms of correlation were recorded between: EPS and EPS Diluted 1.00, Net Profit Margin and Free Cash Flow Margin 0.90, Quick Ratio and Current Ratio 0.87, Return on Total Assets Ratio and Total Asset Turnover Ratio 0.74, Price Earnings-To-Growth Ratio and PE Ratio 0.57. All other variables had correlation values lower than 0.50.

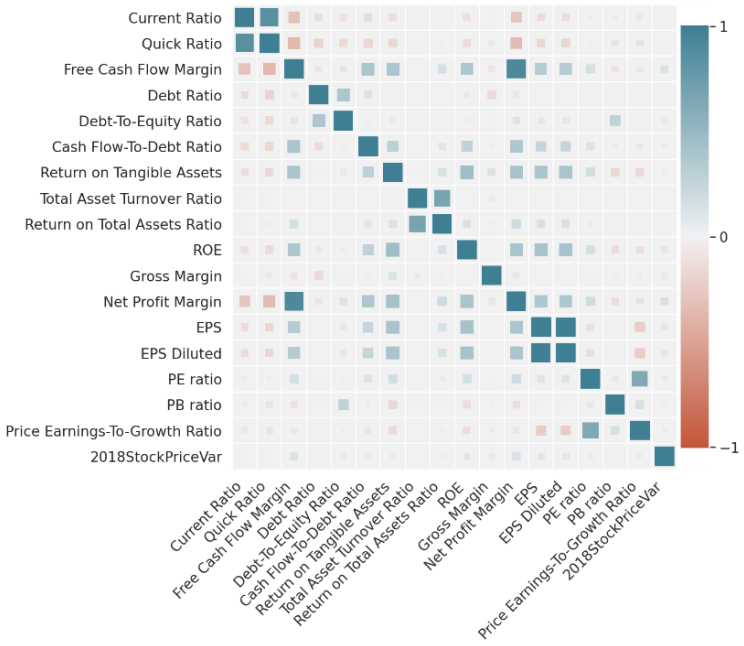


Figure 37. Correlation heatmap analysis, year 2017.

Figure 37 shows the correlation between variables in year 2017 and stock price variance in year 2018. All 17 variables had quite weak correlation with stock price variance. Following results were observed: Current Ratio -0.015, Quick Ratio -0.019, Free Cash Flow Margin 0.12, Debt Ratio -0.021, Debt-To-Equity Ratio -0.076, Cash Flow-To-Debt Ratio 0.066, Return on Tangible Assets 0.02, Total Asset Turnover Ratio 0.0068, Return on Total Assets Ratio 0.039, ROE 0.093, Gross Margin 0.069, Net Profit Margin 0.14, EPS 0.092, EPS Diluted 0.093, PE Ratio -0.063, PB Ratio 0.042. Price Earnings-To-Growth Ratio 0.044. The strongest few positive relationships in terms of correlation were recorded between: EPS and EPS Diluted 1.00, Net Profit Margin and Free Cash Flow Margin 0.92, Quick Ratio and Current Ratio 0.87, Return on Total Assets Ratio and Total Asset Turnover Ratio 0.68, Price Earnings-To-Growth Ratio and PE Ratio 0.63. All other variables had correlation values lower than 0.50.

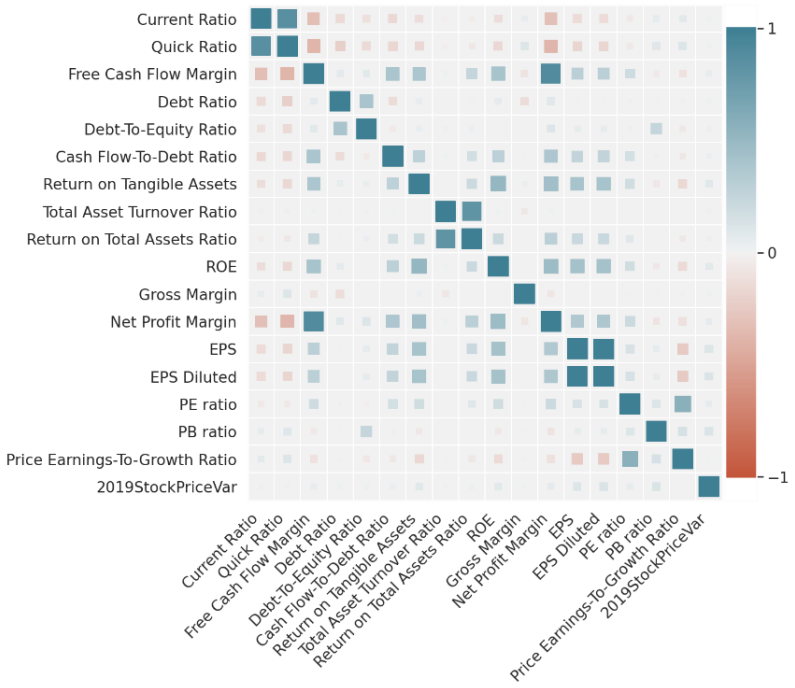


Figure 38. Correlation heatmap analysis, year 2018.

Figure 38 shows the correlation between variables in year 2018 and stock price variance in year 2019. All 17 variables had quite weak correlation with stock price variance. Following results were observed: Current Ratio 0.026, Quick Ratio 0.024, Free Cash Flow Margin 0.055, Debt Ratio 0.027, Debt-To-Equity Ratio 0.031, Cash Flow-To-Debt Ratio 0.038, Return on Tangible Assets 0.045, Total Asset Turnover Ratio -0.017, Return on Total Assets Ratio 0.029, ROE 0.083, Gross Margin 0.021, Net Profit Margin 0.072, EPS 0.12, EPS Diluted 0.12, PE Ratio 0.031, PB Ratio 0.12. Price Earnings-To-Growth Ratio 0.015. The strongest few positive relationships in terms of correlation were recorded between: EPS and EPS Diluted 1.00, Net Profit Margin and Free Cash Flow Margin 0.90, Quick Ratio and Current Ratio 0.86, Return on Total Assets Ratio and Total Asset Turnover Ratio 0.82, Price Earnings-To-Growth Ratio and PE Ratio 0.58. All other variables had correlation values lower than 0.50. These findings are generally consistent with those of the previous scientific literature. It can also be said that the correlation relationships are more or less consistent throughout the years, between the variables.

All in all, after using the linear regression model, it was determined that it was insufficiently accurate for this particular dataset, and hence was abandoned. Correlation research gave a broad perspective of the relationship between financial factors and also the subsequent year's stock price movement. There are at least a few strong relationships, including EPS and EPS Diluted of 1.00, Net Profit Margin and Free Cash Flow Margin of 0.92, Quick Ratio and Current Ratio of 0.87, Return on Total Assets Ratio and Total Asset Turnover Ratio of 0.68, and Price Earnings-to-Growth Ratio and PE Ratio of 0.63. Correlations between stock prices and financial

indicators were found to be poor. As a result, a single financial measure is unlikely to be a good predictor of stock price change, or even to be highly connected with it.

3.3. Classification modeling

The constant advancement of Artificial Intelligence and Machine Learning, together with the increasing accessibility of financial market information to a broader range of investors, has resulted in the emergence of complex trading algorithms that are beginning to have a substantial effect on market dynamics.

As per Efficient Market Hypothesis, a stock market is almost unpredictable. This is because it is hard to outperform the market since share prices already include all relevant information, including prior prices and trade volumes. As a result, price changes are random and do not track any pattern, preventing investors from generating above-average yields without incurring numerous risks (Nobre and Neves, 2019).

This paper's primary focus is the use of XGBoost library to categorize the following year's market direction. Although XGBoost has been used to a wide variety of classification and regression studies, its employment in the US stock market is very limited still. Following figure shows the import of XGBoost Classifier, which was used to answer the main research questions of this paper.

```
from sklearn.metrics import accuracy_score
from xgboost import XGBClassifier
from sklearn.model_selection import cross_val_score
```

Figure 39. Importing the classifier.

Same as with Linear Regression model, here some checks of the model are also valuable information before analyzing the results. Figure 40 shows the preliminary accuracy of the model using Sklearn machine learning accuracy score function. The returned accuracy was 55.7%, which is more than 50%, and that is a first good sign.

```
df_pred = [majority_class] * len(df2[target])
accuracy_score(df2[target], df_pred)
```

```
0.5569866920152091
```

Figure 40. Model accuracy.

After stratifying and building the model the pipeline class was used. Sklearn's pipeline class makes it able to stick few processes into a single estimator. The purpose of this pipeline is

to assemble or put together multiple steps that could be later validated and checked together, while attaching different attributed and parameters. Figure 41 shows the results of the XGBoost model using the pipeline function. Accuracy of train set was 98.51% and accuracy of test set was 64.66%, which is another very good sign of the model. Accuracies are relatively good and this model can be used to analyze results and make conclusions.

```
print(f'Accuracy of train set: {pipeline.score(X_train, y_train)}')
print(f'Accuracy of test set: {pipeline.score(X_test, y_test)}')
```

```
Accuracy of train set: 0.9851473384030418
Accuracy of test set: 0.6466254752851711
```

Figure 41. Accuracies of train and test sets.

Next evaluation step was confusion matrix. It is a table that summarizes the outcomes of a classification study's forecasts. A confusion matrix is a table that summarizes the outcomes of a classification problem's prediction. The amount of right and inaccurate predictions is summarized and divided down by class using counted values. The confusion matrix demonstrates the many scenarios in which the classification model becomes puzzled when making predictions. It provides insight into not just the mistakes produced by the classifier, but also the kind of mistakes produced. This breakdown solves the drawback of relying only on classification accuracy. Figure 42 presents the confusion matrix.

```
Confusion Matrix of test set:
 [[1030  834]
 [ 653 1691]]


Classification report of test set:
              precision    recall  f1-score   support

           0       0.61      0.55      0.58      1864
           1       0.67      0.72      0.69      2344

    accuracy                           0.65      4208
   macro avg       0.64      0.64      0.64      4208
weighted avg       0.64      0.65      0.64      4208
```

Figure 42. Confusion Matrix.

Confusion matrix's values are mostly made of the prediction matrix, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Classification accuracy is calculated by summing TP and TN, then dividing them by the sum of TP, TN, FP, and FN. Recall is calculated by TP being divided by sum of TP and FN. Precision is calculated by TP being divided by the sum of TP and FP. Then F-score is calculated by the

following formula: multiply recall by 2 and then multiply that by Precision, which then is divided by the sum of Recall and Precision. The values are relatively good, model can be used further.

Figure 43 shows the detailed prediction probabilities for classes 0 and 1, which is price goes down or up next year. Results for research purposes are shown only for first 3 and last 3 observations.

```
pipeline.predict_proba(X_test)
```

```
array([[0.22670788, 0.7732921 ],
       [0.3083282 , 0.6916718 ],
       [0.64923704, 0.35076296],
       ...,
       [0.10917717, 0.8908228 ],
       [0.8422121 , 0.15778793],
       [0.35490716, 0.64509284]], dtype=float32)
```

Figure 43. Detailed prediction probabilities.

Figure 44 shows classification model's prediction as an integer, either 0 or 1, also for the same first 3 and last 3 observations.

```
pipeline.predict(X_test)
```

```
array([1, 1, 0, ..., 1, 0, 1])
```

Figure 44. Predictions as integers.

The receiver operating characteristic (ROC) curve is a graphical representation of a binary classifier product's performance. This distinguishing property makes it one of the most widely utilized techniques in a variety of scientific domains. It was first created during World War II to determine if a signal on a radar screen constituted a noise or an object, and is now extensively utilized in medical, radiography, biometrics, bioinformatics, and a variety of other machine learning tasks. Measuring performance is critical in Machine Learning. Thus, while dealing with a classification issue, we may rely on the AUC-ROC Curve. When there is a need to evaluate or depict the performance of a classification issue, we should utilize the receiver operating characteristic (ROC) curve. It is a critical assessment statistic for determining the success of any classification model. (Goksuluk et al., 2016). Figure 45 shows this model's AUC ROC score, which was 0.71. It is an average score, which means model is quite good in distinguishing between the two classes.

```
from sklearn.metrics import roc_auc_score
roc_auc_score(y_test, y_pred_proba)
```

```
0.7064842983272056
```

Figure 45. AUC ROC score.

Figure 46 shows the ROC curve plotted. Red line is represented by ROC and AUC is all the area below it.

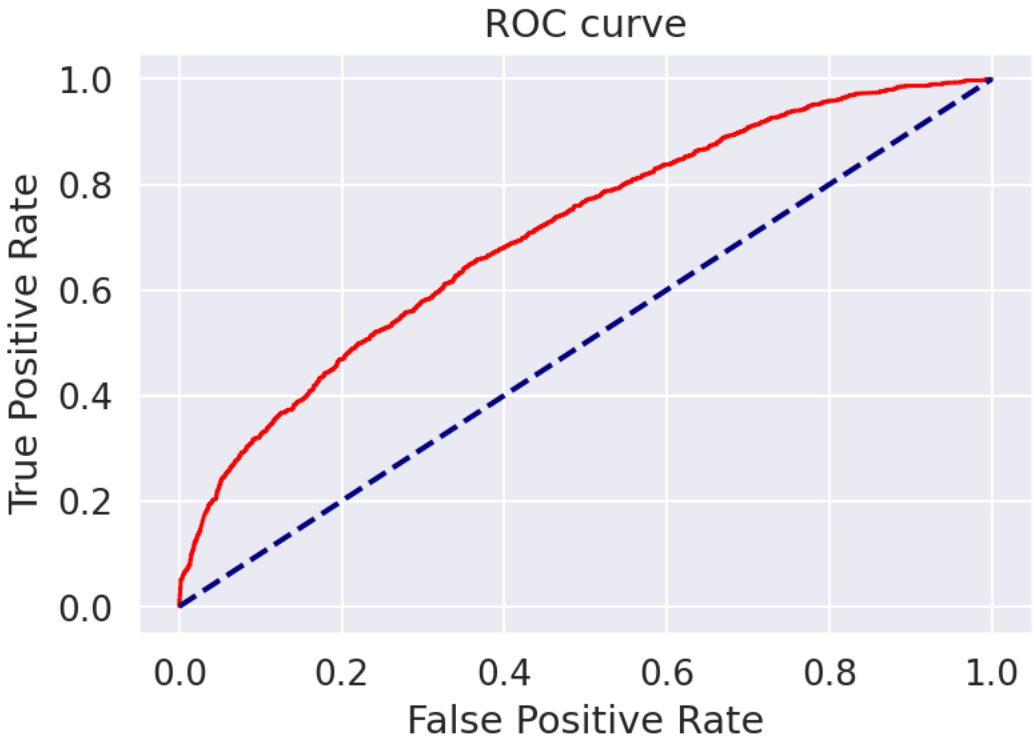


Figure 46. ROC curve.

After evaluating the model and concluding that it is good enough to continue with employing for analysis of the data set, the next step is to actually run it and analyze the results.

The results are presented from three scenarios. First is with 10 iterations, second scenario with 100 iterations and thirst with 1000 iterations. Following figures shows the model attributes and feature importance scores together with standard deviation. Feature importance scores were analyzed and compared to previous literature.

The term "feature importance" refers to the sensitivity analysis technique used in machine learning for determining the link between the inputs and outputs of a trained prediction model. Its objective is to ascertain the independent variables' relative relevance scores in relation to the dependent variable. Notably, they refer only to the relative value of each information in prediction, not to forecast accuracy (Delen et al., 2013).

```
treemodel = XGBClassifier(n_estimators=10, random_state=42, learning_rate=0.1, n_jobs=-1)
treemodel.fit(X_train_transformed, y_train)
```

Weight	Feature
0.0136 ± 0.0052	Return on Tangible Assets
0.0106 ± 0.0067	Free Cash Flow Margin
0.0084 ± 0.0074	Net Profit Margin
0.0058 ± 0.0046	PE ratio
0.0050 ± 0.0024	Price Earnings-To-Growth Ratio
0.0038 ± 0.0051	Return on Total Assets Ratio
0.0023 ± 0.0075	ROE
0.0020 ± 0.0092	Total Asset Turnover Ratio
0.0017 ± 0.0022	EPS
0.0015 ± 0.0037	Cash Flow-To-Debt Ratio
0.0010 ± 0.0022	Current Ratio
0.0007 ± 0.0026	Debt-To-Equity Ratio
-0.0007 ± 0.0013	Debt Ratio
-0.0010 ± 0.0041	Quick Ratio
-0.0011 ± 0.0023	PB ratio
-0.0019 ± 0.0030	EPS Diluted
-0.0026 ± 0.0018	Gross Margin

Figure 47. Model with 10 iterations.

Figure 47 shows the feature importance scores for model with 10 iterations. Few features had negative scores and most of the features had very low scores, which clearly indicates that 10 iterations are just too low and not enough to provide an accountable result.

Next figure presents the attributes for model with 100 iterations and then feature importance scores.

```
treemodel = XGBClassifier(n_estimators=100, random_state=42, learning_rate=0.1, n_jobs=-1)
treemodel.fit(X_train_transformed, y_train)
```

Weight	Feature
0.0173 ± 0.0114	Free Cash Flow Margin
0.0164 ± 0.0061	PE ratio
0.0114 ± 0.0045	Cash Flow-To-Debt Ratio
0.0106 ± 0.0032	Price Earnings-To-Growth Ratio
0.0060 ± 0.0047	EPS Diluted
0.0053 ± 0.0028	Return on Total Assets Ratio
0.0051 ± 0.0032	ROE
0.0048 ± 0.0024	Current Ratio
0.0044 ± 0.0055	EPS
0.0037 ± 0.0079	Total Asset Turnover Ratio
0.0032 ± 0.0021	Debt-To-Equity Ratio
0.0031 ± 0.0082	Return on Tangible Assets
0.0030 ± 0.0060	PB ratio
0.0022 ± 0.0064	Net Profit Margin
0.0022 ± 0.0030	Quick Ratio
0.0019 ± 0.0047	Debt Ratio
0.0005 ± 0.0027	Gross Margin

Figure 48. Model with 100 iterations.

The feature importance scores looked more reasonable, however with more iterations means that the model would have more chance to fix previous models using gradient boosting technique. According to the literature the optimal number of iterations would be 500 to 1000. Number of iterations also depends on physical machine that its ran on, because of the computational resources that the machine learning model requires. Physical machines attributes can be a limitation if too many iterations are chosen.

For research purposes it was chosen to have 1000 iteration model as the final one and analyze its results.

```
treemodel = XGBClassifier(n_estimators=1000, random_state=42, learning_rate=0.1, n_jobs=-1)
treemodel.fit(X_train_transformed, y_train)
```

Weight	Feature
0.0587 ± 0.0087	ROE
0.0468 ± 0.0079	PB ratio
0.0467 ± 0.0053	Price Earnings-To-Growth Ratio
0.0354 ± 0.0033	Net Profit Margin
0.0242 ± 0.0079	EPS
0.0100 ± 0.0049	PE ratio
0.0078 ± 0.0065	Cash Flow-To-Debt Ratio
0.0053 ± 0.0061	Return on Total Assets Ratio
0.0042 ± 0.0029	EPS Diluted
0.0034 ± 0.0067	Total Asset Turnover Ratio
0.0034 ± 0.0047	Current Ratio
0.0033 ± 0.0031	Debt-To-Equity Ratio
0.0023 ± 0.0047	Gross Margin
0.0019 ± 0.0057	Free Cash Flow Margin
0.0013 ± 0.0044	Debt Ratio
0.0009 ± 0.0079	Return on Tangible Assets
0.0006 ± 0.0044	Quick Ratio

Figure 49. Model with 1000 iterations.

According to the XGBoost machine learning model the top 5 most important financial ratios to predict next year's stock price is ROE with feature importance score of 0.0587 and standard deviation of 0.0087, PB Ratio 0.0468 and standard deviation 0.0079, Price Earnings-To-Growth Ratio 0.0467 and standard deviation 0.0053, Net Profit Margin 0.0354 and standard deviation 0.0033, and EPS 0.0242 and standard deviation 0.0079.

Weight	Feature
0.2291	ROE
0.1827	PB ratio
0.1823	Price Earnings-To-Growth Ratio
0.1382	Net Profit Margin
0.0945	EPS
0.0390	PE ratio
0.0304	Cash Flow-To-Debt Ratio
0.0207	Return on Total Assets Ratio
0.0164	EPS Diluted
0.0133	Total Asset Turnover Ratio
0.0133	Current Ratio
0.0129	Debt-To-Equity Ratio
0.0090	Gross Margin
0.0074	Free Cash Flow Margin
0.0051	Debt Ratio
0.0035	Return on Tangible Assets
0.0023	Quick Ratio

Figure 50. Standardized weight of features.

Figure 50 shows standardized weight of the features. Top 5 most significant features according to model hold following weights: ROE 0.2291, PB Ratio 0.1827, Price Earnings-To-Growth Ratio 0.1823, Net Profit Margin 0.1382, EPS 0.0945. These results correlate with some of the previous literature, specifically with the works from Arkan (2016), Martani et al. (2009), and Petcharabul and Romprasert's (2014) study.

Machine learning models tend to be "black boxes" and how the results are calculated or approached are rarely explained or even looked at. Following figure with the help of SHAP function shows how a XGBoost machine learning model perceives and evaluates a single observation.

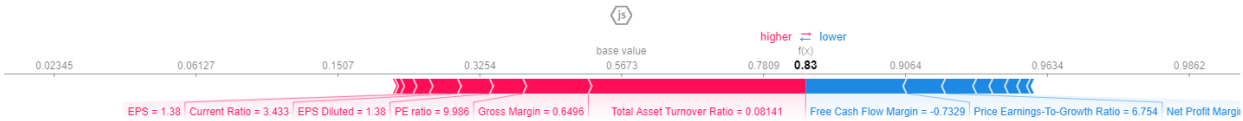


Figure 51. SHAP values.

Red features push the prediction closer to 1, meaning next year the stock price will go up and blue features push the prediction closer to 0, meaning next year stock price will go down. Base value was 0.5673 and particularly for this observation the predicted output is 0.83, meaning the model is leaning towards the price going up next year.

If previous figure explained model's approach for a single observation, the following figure shows relationship for all observations used in the data set.

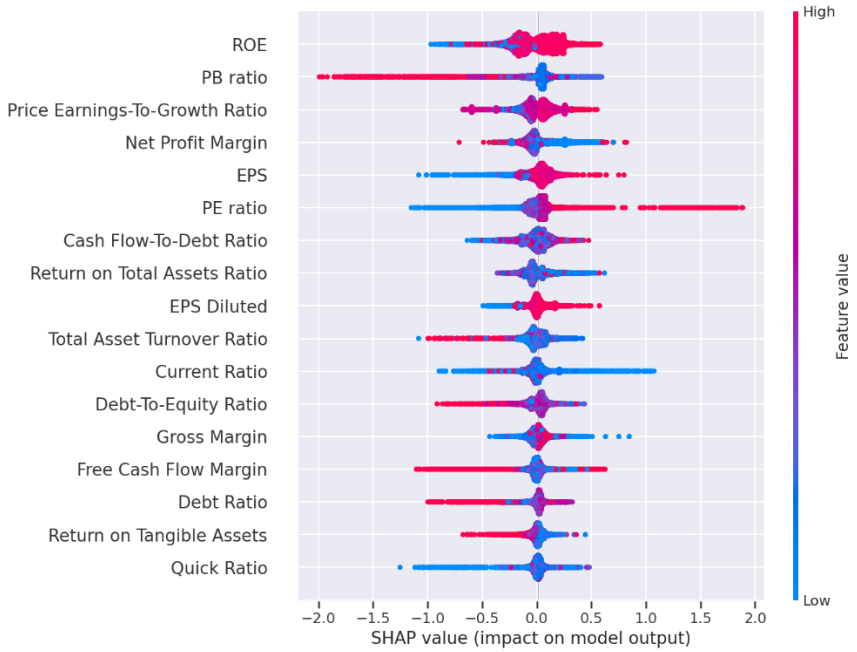


Figure 52. SHAP value relationships.

X axis holds values for impact on the model output. Please note, that these values are not probabilistic, but logarithmic. Red color means feature had a high value and blue color means feature had a low value in the observation. According to the distribution it can be seen how big the impact was. Financial ratios are ordered by the importance, top being the most important according to the machine learning model. Overall, the situation looks like there was no financial ratio which really was different from all other in terms of impact. However, it can be seen that if ROE is trending higher, the model assumes that the price will increase of the stock. Also, model suggests that high PB ratio can be an indication for stock price going lower next year, in other words stock being overvalued. Interestingly, for PE ratio the model assumes that high PE ratio might suggest price increase next year. Generally, it may be of the situation that high PE ratio would signal that those investors may pay a higher price for the share today, because of high growth expectations.

To summarize the classification modelling results, the XGBoost machine learning model performed admirably. After 1000 iterations of extreme gradient boosting, the machine learning model determined that the top five most important financial ratios for predicting the following year's stock price were ROE with feature importance score of 0.0587 and standard deviation of 0.0087, PB Ratio 0.0468 and standard deviation 0.0079, Price Earnings-To-Growth Ratio 0.0467 and standard deviation 0.0053, Net Profit Margin 0.0354 and standard deviation 0.0033, and EPS 0.0242 and standard deviation 0.0079. After standardizing the weights assigned to the features, the following weights were assigned to the model's top five most significant features: ROE 0.2291; PB Ratio 0.1827; Price Earnings-to-Growth Ratio 0.1823; Net Profit Margin 0.1382; and EPS

0.0945. These findings corroborate with prior research, most notably Arkan (2016), Martani et al. (2009), and Petcharabul and Romprasert's (2014) study, in terms of which features are most significant to the algorithm. However, the feature importance scores are not high, meaning that those financial indicators may not have a significant impact or high predictive power for next years' stock price movement.

CONCLUSIONS AND PROPOSALS

Shareholders and management are usually concerned with maintaining and improving the financial health of publicly listed organizations. Investors seek to optimize their earnings by distributing their excess funds to lucrative businesses in the expectation of achieving long-term returns on respective holdings. Along with many other financial indicators, share price is a financial ratio upon which traders pay special attention when making information-based investing choices. Additionally, shareholders must be aware of the financial factors that determine expected performance.

The purpose of this study was to examine the relationship between financial ratios and share price returns of next year in the U.S. stock market in order to get a better understanding of the relationship. The important financial measures that have a substantial correlation with share prices were identified quantitatively. Given the current artificial intelligence excitement regarding the case of financial services, the purpose was to employ machine learning model for forecasting share price direction using a set of financial measures.

The first research question was expressed as follows: “What relationship exists between financial ratios and stock market performance in the United States of America from 2014 to 2018?”. The subject was investigated via a review of many academic publications and the application of machine learning model to a large sample of US stock market information. At the moment, there is no agreement in the research about which financial parameters have the greatest influence on returns in international markets. Nevertheless, the academic literature assessment indicated that profitability and market value ratios seemed to have a greater relationship with share price directions compared to other types of ratios.

The academic literature review in-part addressed the second question: “Which of the financial ratios possesses stronger relationship with stock price performance?”. According to the literature assessment, the profitability ratios had the highest correlation with the share price direction. This validated the use of profitability ratios in the quantitative part of the study.

In the empirical part, distribution of stock price variances was first examined. It can be said that the price variances of all 5 years were normally distributed with positive or negative leaning based on that year’s performance. No anomalies in terms of distribution were observed. Stock price variance analysis based on sectors showed that sectors do react and move slightly differently. It can also be distinguished that there are so called “defensive” sectors, where stock prices are not diving down when all other sectors are. However, the “defensive” sectors are not going to provide large returns if looking on average stock price distribution. Then, visualized relationships revealed that the healthcare and technology sectors dominate in terms of high growth and high risk. It can be assumed that the reason for this is the number of new companies or start-

ups and their potential in these two sectors. After application of linear regression model, it proved to be not so accurate for this particular dataset, therefore it was abandoned. Correlation analysis provided with the general view of relatability between the financial variables and also the stock price movement of following year. It was found that there are at least a few strong relationships, such as EPS and EPS Diluted 1.00, Net Profit Margin and Free Cash Flow Margin 0.92, Quick Ratio and Current Ratio 0.87, Return on Total Assets Ratio and Total Asset Turnover Ratio 0.68, Price Earnings-To-Growth Ratio and PE Ratio 0.63. When it came to the stock price and financial variables, the correlation proved to be weak. Therefore, it can be interpreted that a single financial ratio is unlikely to be a strong predictor of stock price movement, or at least not strongly correlated. When it came to classification modelling, the machine learning model XGBoost proved to be of reasonable accuracy. After 1000 iterations of extreme gradient boosting, according to the machine learning model **the top 5 most important financial ratios to predict the following year's stock price were ROE with feature importance score of 0.0587 and standard deviation of 0.0087, PB Ratio 0.0468 and standard deviation 0.0079, Price Earnings-To-Growth Ratio 0.0467 and standard deviation 0.0053, Net Profit Margin 0.0354 and standard deviation 0.0033, and EPS 0.0242 and standard deviation 0.0079.** After standardizing the weights of the features these were following weights of top 5 most significant features according to the model: ROE 0.2291, PB Ratio 0.1827, Price Earnings-To-Growth Ratio 0.1823, Net Profit Margin 0.1382, EPS 0.0945. These results correlate with some of the previous literature, specifically with the works from Arkan (2016), Martani et al. (2009), and Petcharabul and Romprasert's (2014) study.

In summary of the all-empirical findings, it can be said that studied **financial ratios do possess a relationship with the stock market performance, however not a strong one**. **Results showed that standalone financial ratios will not provide sufficient material to forecast movement of the stock price.** However, if used all together, the picture can be clearer, and if combined with additional facts, like micro and macro-economic environment and news articles, the forecast could be improved. Also, this paper demonstrates the appropriateness of machine learning approaches for determining the relationship between financial measures and stock returns in the US market. While the model's forecasting accuracy limits are important, the results showing the meaningful financial ratios corroborate the existing literature.

However, there were some limitations for the study, which could have held back the results of this paper. The model's accuracy could have been much better still. The not so high accuracy could be explained by data used, particularly the size and quality. The larger the dataset, the more training and learning material the machine learning model has. Further research employing many more years could reduce this limitation. Further, more financial indicators could be used and also other real-world indicators such as news articles or even tweets could be employed. Also, other

models could be used in terms of comparison or even combination to arrive at more accurate predictions.

LIST OF REFERENCES AND SOURCES

Alsagheer, R., Alharan, A., Al-Haboobi, A. (2017). *Popular Decision Tree Algorithms of Data Mining Techniques: A Review.* Viewed on 2021-04-18. Internet access: https://www.researchgate.net/publication/317731072_Popular_Decision_Tree_Algorithms_of_Data_Mining_Techniques_A_Review

Anwaar, M. (2016). *Impact of Firms' Performance on Stock Returns (Evidence from Listed Companies of FTSE-100 Index London, UK).* Viewed on 2021-02-20. Internet access: https://globaljournals.org/item/5870-impact-of-firms-performance-on-stock-returns-evidence-from-listed-companies-of-ftse-100-index-london-uk

Atsalakis, G. S., & Valavanis, K. P. (2009). *Surveying stock market forecasting techniques – Part II: Soft computing methods.* Expert Systems with Applications. Viewed on 2021-02-20. Internet access: https://doi.org/10.1016/j.eswa.2008.07.006

Arkan, T. (2016). *The Importance of Financial Ratios in Predicting Stock Price Trends: A Case Study in Emerging Markets.* Viewed on 2021-03-27. Internet access: https://www.ceeol.com/search/article-detail?id=623339

Beaver, W. H. (1968). *Market Prices, Financial Ratios, and the Prediction of Failure*. Journal of Accounting Research, 6. Viewed on 2021-04-17. Internet access: https://doi.org/10.2307/2490233

Casalicchio, G., Molnar, C., Bischl, B. (2019). *Visualizing the Feature Importance for Black Box Models.* Viewed on 2021-04-10. Internet access: https://link.springer.com/chapter/10.1007/978-3-030-10925-7_40

Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., & Nobrega, J. P. (2016). *Computational Intelligence and Financial Markets: A Survey and Future Directions*. Expert Systems with Applications, 55. Viewed on 2021-02-19. Internet access: http://dx.doi.org/10.1016/j.eswa.2016.02.006

Chairakwattana, K., & Nathaphan, S. (2014). Stock Return Predictability by Bayesian Model Averaging: Evidence from Stock Exchange of Thailand. Viewed on 2021-03-13. Internet access: http://dx.doi.org/10.13140/RG.2.1.1161.1683

Chan, Y. (2004). *Biostatistics 201: linear regression analysis.* Viewed on 2021-04-18. Internet access: https://pubmed.ncbi.nlm.nih.gov/14985842/

Chen, S.-W., & Shen, C.-H. (2009). *Is the Stock Price Higher than that Implied by the Fundamentals?* International Research Journal of Finance and Economics. Viewed on 2021-02-21. Internet access: https://www.researchgate.net/publication/265290218_Is_the_Stock_Price_Higher_than_that_Implied_by_the_Fundamentals

Chen, T., Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Viewed on 2021-04-18. Internet access: https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

Davydenko, A., Fildes, R. (2013). *Measuring Forecasting Accuracy: Problems and Recommendations (by the Example of SKU-Level Judgmental Adjustments).* Viewed on 2021-04-10. Internet access: http://dx.doi.org/10.1007/978-3-642-39869-8_4

Delen, D., Kuzey, C., & Uyar, A. (2013). *Measuring firm performance using financial ratios: A decision tree approach.* Expert Systems with Applications. Viewed on 2021-02-20. Internet access: https://doi.org/10.1016/j.eswa.2013.01.012

Dey, A. (2016). *Machine Learning Algorithms: A Review.* International Journal of Computer Science and Information Technologies, Vol. 7. Viewed on 2021-04-18. Internet access: https://pdf4pro.com/view/machine-learning-algorithms-a-review-ijcsit-4d28d1.html

Emamgholipour, M., Pouraghajan, A., Tabari, N.A., Milad, Haghparast, & Shirsavar, A. (2013). *The Effects of Performance Evaluation Market Ratios on the Stock Return: Evidence from the Tehran Stock Exchange.* Viewed on 2021-02-20. Internet access: https://www.semanticscholar.org/paper/The-Effects-of-Performance-Evaluation-Market-Ratios-Emamgholipour-Pouraghajan/63162f72855924a5f58d6742308723990b39bd56

Fama, E. F., & French, K. R. (1988). *Dividend yields and expected stock returns.* Journal of Financial Economics. Viewed on 2021-03-11. Internet access: https://doi.org/10.1016/0304-405X(88)90020-7

Fernández-Gavilanes, M., Álvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., Javier González-Castaño, F. (2016). *Unsupervised method for sentiment analysis in online texts.* Expert Systems with Applications, 58. Viewed on 2021-04-18. Internet access: https://doi.org/10.1016/j.eswa.2016.03.031

Flach, P. (2016). *ROC Analysis*. Viewed on 2021-04-10. Internet access: https://doi.org/10.1007/978-1-4899-7502-7_739-1

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow (1st ed.).* Viewed on 2021-03-13.

Gharehchopogh, F. (2013). *A linear regression approach to prediction of stock market trading volume: a case study*. International Journal of Managing Value and Supply Chains. Viewed on 2021-04-10. Internet access: https://www.academia.edu/26292464/A_Linear_Regression_Approach_to_Prediction_of_Stock_Market_Trading_Volume_A_Case_Study

Goksuluk, D., Korkmaz, S., Zararsiz, G., Karaagaoglu, A. (2016*). easyROC: An Interactive Web-tool for ROC Curve Analysis Using R Language Environment*. Viewed on 2021-04-10. Internet access: https://journal.r-project.org/archive/2016/RJ-2016-042/RJ-2016-042.pdf

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). *Literature review: Machine learning techniques applied to financial market prediction.* Viewed on 2021-03-13. Internet access: https://doi.org/10.1016/j.eswa.2019.01.012

Heydari, A., Tavakoli, M., Salim, N. (2016). *Detection of fake opinions using time series*. Expert Systems with Applications, 58. Viewed on 2021-04-18. Internet access: https://doi.org/10.1016/j.eswa.2016.03.020

Hooker, S., Erhan, D., Kindermans, P., Kim, B. (2019). *A Benchmark for Interpretability Methods in Deep*

Neural Networks. Viewed on 2021-04-18. Internet access: https://arxiv.org/pdf/1806.10758.pdf

Huang, W., Mollick, A., Nguyen, K. (2016). *U.S. stock markets and the role of real interest rates*. The Quarterly Review of Economics and Finance, 59. Viewed on 2021-04-10. Internet access: https://doi.org/10.1016/j.qref.2015.07.006

Imandoust, S., Bolandraftar, M. (2014). *Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange.* Viewed on 2021-04-24. Internet access: https://www.researchgate.net/publication/315924718_Forecasting_the_direction_of_stock_market_index_movement_using_three_data_mining_techniques_the_case_of_Tehran_Stock_Exchange

Ketkar, N. (2017). *Deep Learning with Python: A Hands-on Introduction. New York: Apress.* Viewed on 2021-03-13. Internet access: *https://www.academia.edu/42286263/Deep_Learning_with_Python_A_Hands_on_Introduction_Nikhil_Ketkar?auto=download*

Kheradyar, S., Ibrahim, I., & Mat Nor, F. (2011). Stock Return Predictability with Financial Ratios. International Journal of Trade, Economics and Finance. Viewed on 2021-03-13. Internet access: http://www.ijtef.org/papers/137-S00072.pdf

Kotsiantis, S. (2007). *Supervised Machine Learning: A Review of Classification Techniques.* Viewed on 2021-04-18. Internet access: https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Kim, S. Y., & Upneja, A. (2014). Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Viewed on 2021-03-13. Internet access: https://doi.org/10.1016/j.econmod.2013.10.005

Kumari, K., Yadav, S. (2018). *Linear regression analysis study.* Journal of the Practice of Cardiovascular Sciences 4. Viewed on 2021-04-18. Internet access: https://www.researchgate.net/publication/324944461_Linear_regression_analysis_study

Lewellen, J. (2004). *Predicting returns with financial ratios.* Viewed on 2021-03-13. Internet access: https://doi.org/10.1016/j.jfineco.2002.11.002

Ma, L., Ausloos, M., Schinckus, C., & Chong, F. (2018). *Fundamental Analysis in China: An Empirical Study of the Relationship between Financial Ratios and Stock Prices.* Theoretical Economics Letters. Viewed on 2021-02-21. Internet access: http://dx.doi.org/10.4236/tel.2018.815209

Martani, D., Mulyono, Khairurizka, R. (2009). *The effect of financial ratios, firm size, and cash flow from operating activities in the interim report to the stock return.* Viewed on 2021-03-27. Internet access: http://www.davidpublisher.com/Public/uploads/Contribute/556429e91247b.pdf

Mitchell., T. (2006). *The Discipline of Machine Learning.* Viewed on 2021-03-27. Internet access: http://ra.adm.cs.cmu.edu/anon/usr0/ftp/anon/ml/CMU-ML-06-108.pdf

Musallam, S. R. (2018). *Exploring the Relationship between Financial Ratios and Market Stock Returns.* Eurasian Journal of Business and Economics. Viewed on 2021-03-13. Internet access: http://dx.doi.org/10.17015/ejbe.2018.021.06

Nobre, J., Neves, R. (2019). *Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets*. Viewed on 2021-04-10. Internet access: https://doi.org/10.1016/j.eswa.2019.01.083

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. Viewed on 2021-03-13. Internet access: http://dx.doi.org/10.5430/afr.v7n1p109

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). *Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques.* Expert Systems with Applications, 42. Viewed on 2021-02-20. Internet access: https://doi.org/10.1016/j.eswa.2014.07.040

Pech, C. O., Noguera, M., & White, S. (2015). *Financial ratios used by equity analysts in Mexico and stock returns*. Contaduría y Administración, 60. Viewed on 2021-02-20. Internet access: https://doi.org/10.1016/j.cya.2015.02.001

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & et al. (2012). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12. Viewed on 2021-02-20. Internet access: https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python

Petcharabul, P., & Romprasert, S. (2014). *Technology Industry on Financial Ratios and Stock Returns.* Journal of Business and Economics, 5. Viewed on 2021-02-20. Internet access: http://dx.doi.org/10.21511/imfi.17(2).2020.07

Ross, S. A., Westerfield, R. W., & Jaffe, J. F. (2002). *Corporate Finance (6th ed.).* New York: The McGraw-Hill Companies.

Sarker, I., Kayes, A., Watters, P. (2019). *Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage.* Viewed on 2021-04-18. Internet access: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0219-y

Singh, T., Mehta, S., Varsha, M. (2011). *Macroeconomic factor and stock returns: Evidence from Taiwan.* Viewed on 2021-04-18. Internet access: https://www.researchgate.net/publication/228985237_Macroeconomic_factor_and_stock_returns_Evidence_from_Taiwan

Singh, A., Thakur, N., Sharma, A. (2016). *A review of supervised machine learning algorithms.* Viewed on 2021-04-18. Internet access: *https://ieeexplore.ieee.org/abstract/document/7724478/metrics#metrics*

Sun, J., Jia, M.-y., & Li, H. (2011). *AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies.* Expert Systems with Applications, 38. Viewed on 2021-02-20. Internet access: http://dx.doi.org/10.1016/j.eswa.2011.01.042

Tang, J., Alelyani, S., Liu, H. (2014). *Feature Selection for Classification: A Review.* Viewed on 2021-04-18. Internet access: http://www.cvs.edu.in/upload/feature_selection_for_classification.pdf

Thrun, S., Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. (2019). *A guide to deep learning in healthcare.* Viewed on 2021-03-27. Internet access: https://www.nature.com/articles/s41591-018-0316-z

Tsai, C.-F., Lin, Y.-C., Yen, D. C., & Chen, Y.-M. (2011). *Predicting stock returns by classifier ensembles.* Applied Soft Computing, 11. Viewed on 2021-03-27. Internet access: https://doi.org/10.1016/j.asoc.2010.10.001

VanderPlas, J., (2017). *Python Data Science Handbook.* Published by O'Reilly Media, Inc.

Wang, H., Jiang, Y., & Wang, H. (2009). *Stock Return Prediction Based on Bagging-Decision Tree.* Proceedings of 2009 IEEE International Conference on Grey Systems and Intelligent Services. Viewed on 2021-03-28. Internet access: https://doi.org/10.1109/GSIS.2009.5408165

Zhang, X., Chen, W. (2019). *Stock Selection Based on Extreme Gradient Boosting.* Viewed on 2021-03-27. Internet access: https://doi.org/10.23919/ChiCC.2019.8865781

Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). *Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction*. Expert Systems with Applications, 58. Viewed on 2021-03-27. Internet access: https://www.ii.pwr.edu.pl/~tomczak/PDF/%5BMZSTJT%5D.pdf

Zhou, Z., Hooker, G. (2020). *Unbiased Measurement of Feature Importance in Tree-Based Methods.* Viewed on 2021-04-18. Internet access: https://arxiv.org/pdf/1903.05179.pdf

Zhou, Z. (2009). *Ensemble Learning.* Viewed on 2021-04-18. Internet access: https://doi.org/10.1007/978-0-387-73003-5_293

SUMMARY IN ENGLISH

THE RELATIONSHIP BETWEEN STOCK PERFORMANCE AND FINANCIAL RATIOS: AN APPROACH BASED ON MACHINE LEARNING

Martynas GINTALAS

Master Thesis

Finance and Banking Master Programme

Faculty of Economics and Business Administration, Vilnius University

Supervisor prof. Algimantas Laurinavičius, Vilnius, 2022

SUMMARY

75 pages, 2 tables, 52 figures, 57 references.

The intention of this study is to examine the relationship between financial ratios and stock performance in the U.S. stock markets between the years of 2014 and 2018. It reviews the existing literature on financial ratios and their application to financial results prediction, as well as machine learning in the field of stock market prediction. Besides that, the aim is to develop and apply machine learning model for share price prediction that would incorporate a variety of financial ratios as parameters.

Master thesis is divided into three sections: literature review, research and findings, and conclusions and suggestions.

This paper is centered on the principles of the relationship between financial ratios and stock market returns. Profitability ratios acting as indicators of company success and returns, and mathematical and machine learning-based prediction methods. A review of the literature seeks to assess which financial ratios have a strong correlation with stock returns, regardless of whether mathematical or machine learning approaches are used. Additionally, prior research data and methodologies are explored.

After the literature analysis, author goes to inspect the relationship between financial ratios and stock returns in the U.S. stock exchanges from 2014 to 2018 using novel machine learning technique and a large dataset. Quantitative techniques were employed to determine which financial ratios have a strong correlation with stock return. Additionally, a machine learning model was used to investigate which ratios could forecast stock returns, utilizing a set of financial ratios that have been shown in the literature to have a strong correlation with company share price results. The empirical part of this master thesis could be summarized in a few clear steps.

The very first phase was data collection, which included obtaining raw data from a third-party service. Followed by workspace creation in Python and developing the dataset. All the empirical portion of the study was done using Python. Next step was exploratory data analysis in order to understand it and prepare for cleaning procedures. After a clear picture of dataset, cleaning was done, where missing values and outliers were handled accordingly. Afterwards, visualizations of relationships and distributions, feature engineering and selection was performed. Finally, prediction modelling using regression and classification methods was performed by employing machine learning algorithm XGBoost. The empirical findings chapter is split into three subsections: overall analysis findings, linear regression and correlation findings, and classification modeling findings using the XGBoost classifier.

Finally, author concluded that the top 5 most important financial ratios to predict the following year's stock price were ROE, PB Ratio, Price Earnings-To-Growth Ratio, Net Profit Margin, and EPS. Moreover, in summary of the all-empirical findings, it can be said that studied financial ratios do possess a relationship with the stock market performance, however not a strong one. Results showed that standalone financial ratios will not provide sufficient material to forecast movement of the stock price.

SUMMARY IN LITHUANIAN

AKCIJŲ RINKOS REZULTATŲ IR FINANSINIŲ RODIKLIŲ SĄNTYKIS: MAŠININIO MOKYMOSI METODAS

Martynas GINTALAS

Magistro baigiamasis darbas

Finansų ir Bankininkystės Magistro Programa

Ekonomikos ir Verslo Administravimo Fakultetas, Vilniaus Universitetas

Darbo vadovas prof. Algimantas Laurinavičius, Vilnius, 2022

SĄNTRAUKA

75 lapai, 2 lentelės, 52 iliustracijos, 57 nuorodos.

Šio tyrimo tikslas – išnagrinėti ryšį tarp finansinių rodiklių ir akcijų rezultatų JAV akcijų rinkose 2014–2018 m. Tyrime apžvelgiama esama literatūra apie finansinius rodiklius ir jų taikymą finansinių rezultatų prognozavimui, taip pat mašininiam mokymuisi akcijų rinkos prognozavimo srityje. Be to, tikslas yra pritaikyti mašininio mokymosi modelį akcijų kainos prognozavimui, kuris kaip parametrus apimtų įvairius finansinius rodiklius.

Magistro baigiamasis darbas suskirstytas į tris dalis: literatūros apžvalga, tyrimai, bei išvados ir pasiūlymai.

Šiame tyrime pagrindinis dėmesys skiriamas finansinių rodiklių ir akcijų rinkos grąžos santykio principams. Naudojami pelningumo koeficientai, veikiantys kaip įmonės sėkmės ir grąžos rodikliai, ir matematiniai bei mašininiu mokymusi pagrįsti prognozavimo metodai. Literatūros apžvalga siekia įvertinti, kurie finansiniai rodikliai turi stiprią koreliaciją su akcijų grąža, nepaisant to, ar naudojami matematiniai, ar mašininio mokymosi metodai. Be to, nagrinėjami ankstesnių tyrimų duomenys ir metodikos.

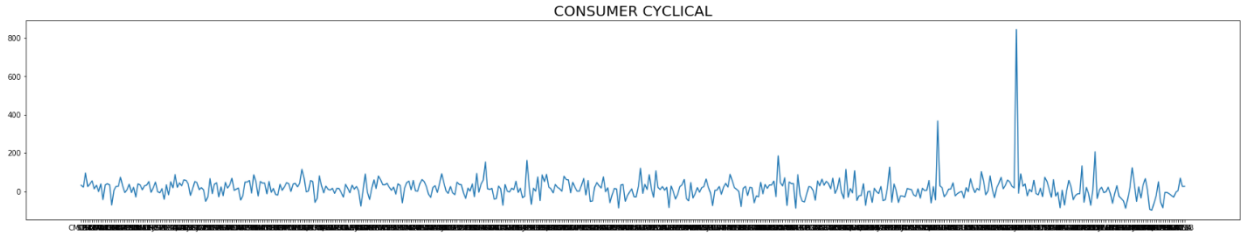
Atlikęs literatūros analizę, autorius imasi nagrinėti finansinių rodiklių ir akcijų grąžos ryšį JAV biržose 2014–2018 m., naudodamas naują mašininio mokymosi techniką ir didelį duomenų rinkinį. Siekiant nustatyti, kurie finansiniai rodikliai stipriai koreliuoja su akcijų grąža, buvo naudojami kiekybiniai metodai. Be to, buvo naudojamas mašininio mokymosi modelis, siekiant ištirti, kurie koeficientai galėtų prognozuoti akcijų grąžą, naudojant finansinių rodiklių rinkinį, kuris, kaip rodo literatūroje, turi tvirtą ryšį su įmonės akcijų kainos rezultatais. Empirinę šio magistro darbo dalį būtų galima apibendrinti keliais aiškiais žingsniais. Pats pirmasis etapas buvo duomenų rinkimas, kuris apėmė neapdorotų duomenų gavimą iš trečiosios šalies paslaugos. Po to

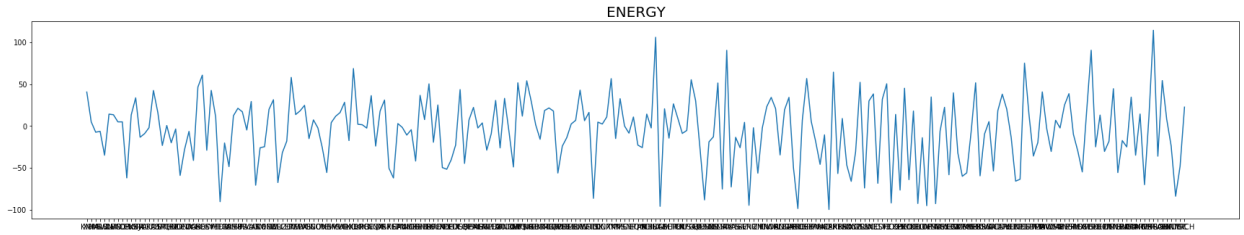
buvo sukurta darbo aplinka Python programa ir kuriamas duomenų rinkinys. Visa empirinė tyrimo dalis buvo atlikta naudojant Python. Kitas žingsnis buvo tiriamoji duomenų analizė, siekiant juos suprasti ir pasiruošti valymo procedūroms. Gavus aiškų duomenų rinkinio vaizdą, buvo atliktas valymas, kuriame atitinkamai buvo tvarkomos trūkstamos reikšmės ir nuokrypiai. Vėliau buvo atliekamos ryšių ir paskirstymo vizualizacijos, funkcijų inž+inerija ir atranka. Galiausiai, prognozavimo modeliavimas naudojant regresijos ir klasifikavimo metodus buvo atliktas naudojant mašininio mokymosi algoritmą “XGBoost”. Empirinių išvadų skyrius yra padalintas į tris poskyrius: bendrosios analizės išvados, tiesinės regresijos ir koreliacijos išvados, bei klasifikavimo modeliavimo išvados naudojant “XGBoost” klasifikatorių.

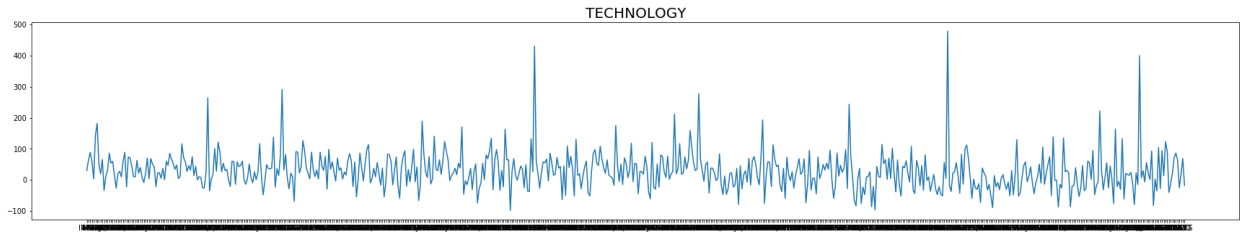
Galiausiai autorius padarė išvadą, kad 5 svarbiausi finansiniai rodikliai, leidžiantys prognozuoti kitų metų akcijų kainą, buvo nuosavybės grąža, PB santykis, kainos pelno ir augimo santykis, grynojo pelno marža ir uždarbis per akciją. Be to, apibendrinant visas empirines išvadas, galima teigti, kad ištirti finansiniai rodikliai turi ryšį su akcijų rinkos rezultatais, tačiau ne stiprų. Rezultatai parodė, kad atskiri finansiniai rodikliai nepateiks pakankamai medžiagos prognozuoti akcijų kainos judėjimą.

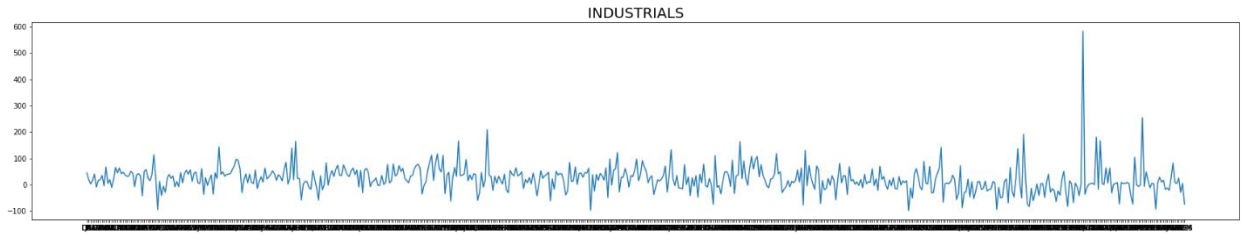
APPENDIX

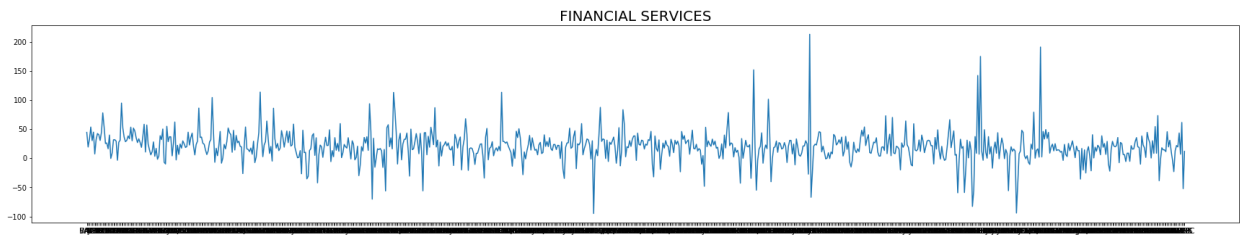
Sectors before cleaning

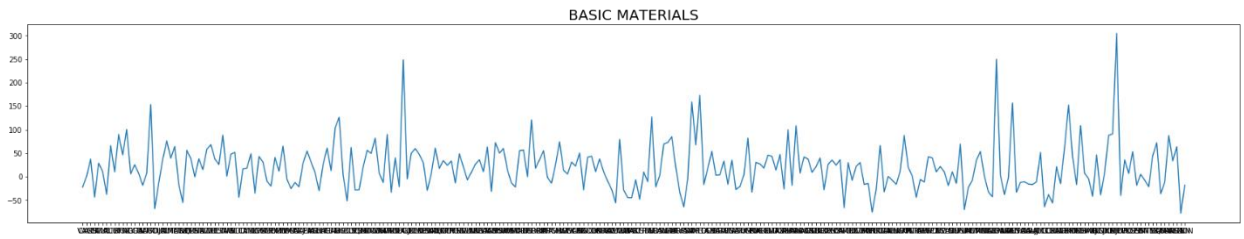


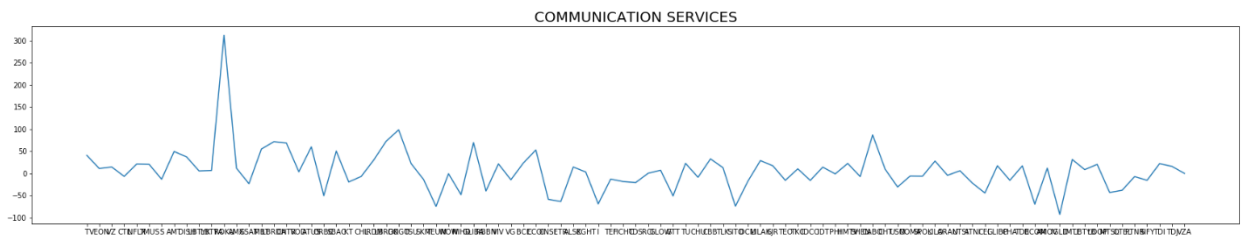


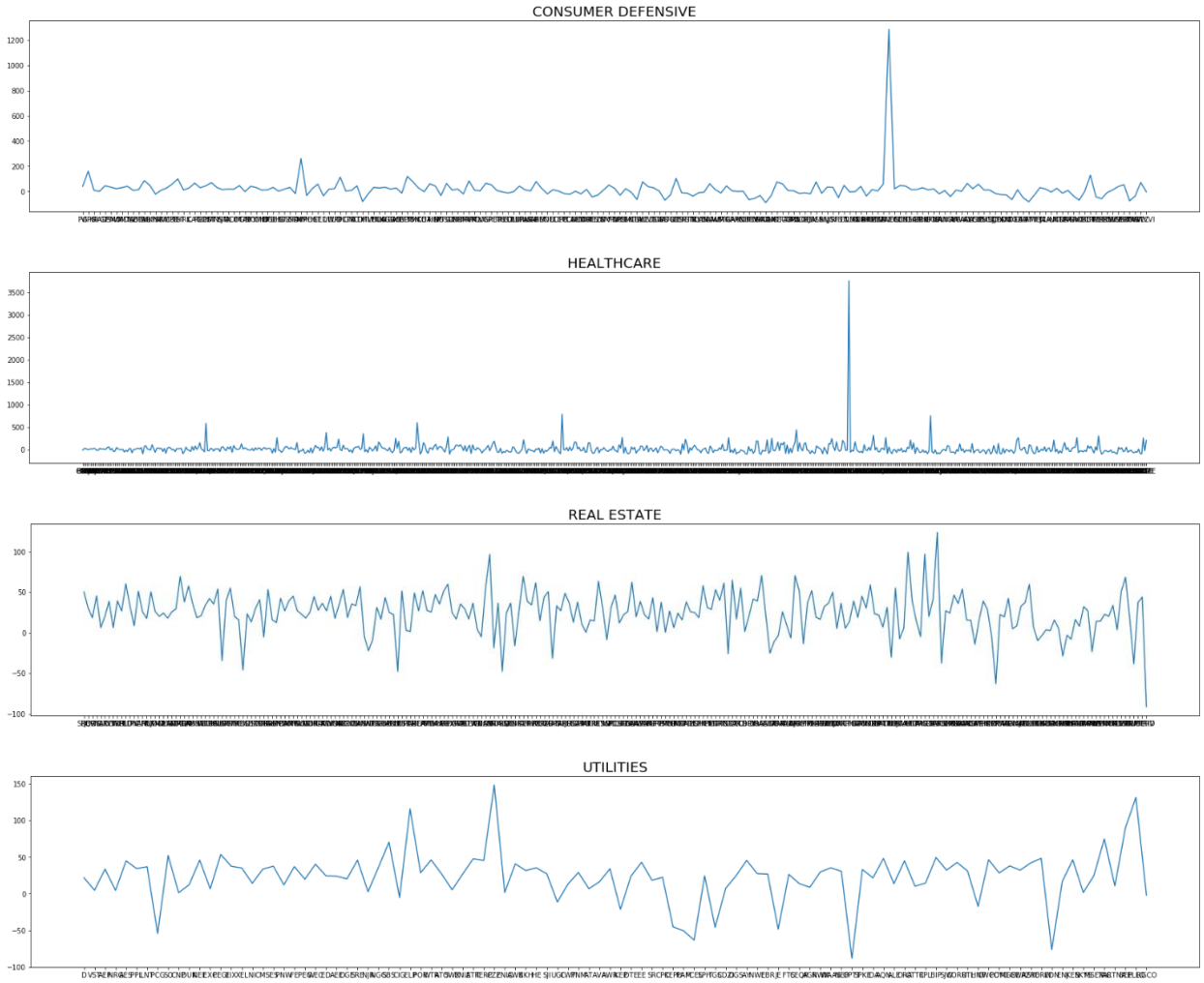




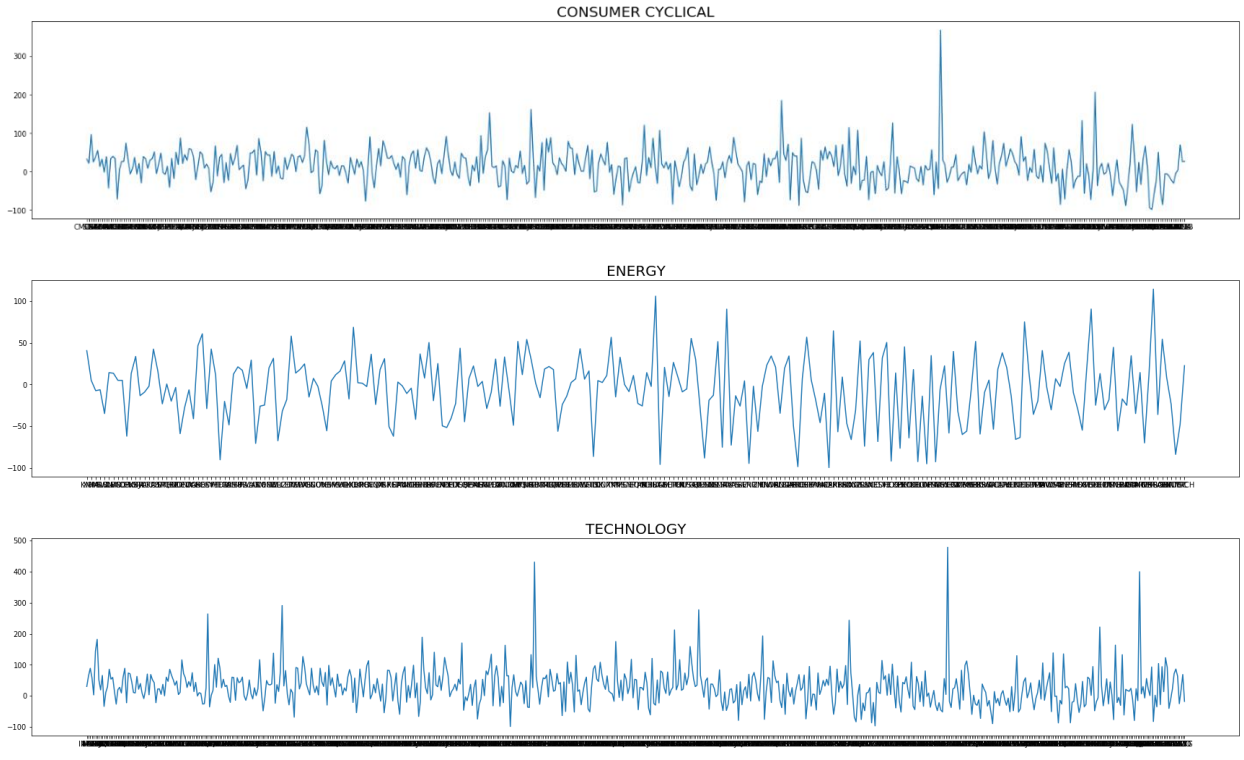


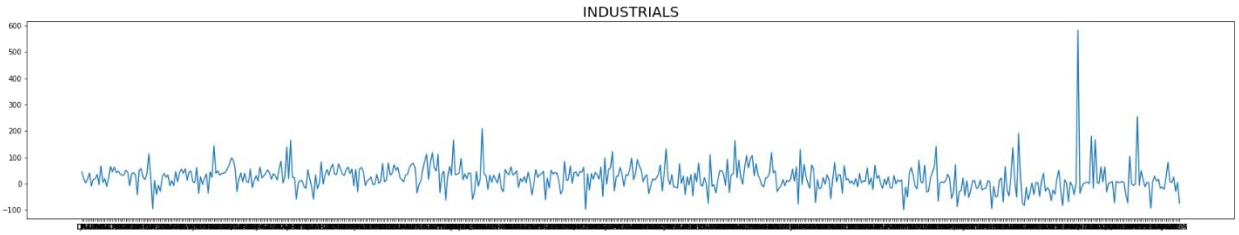


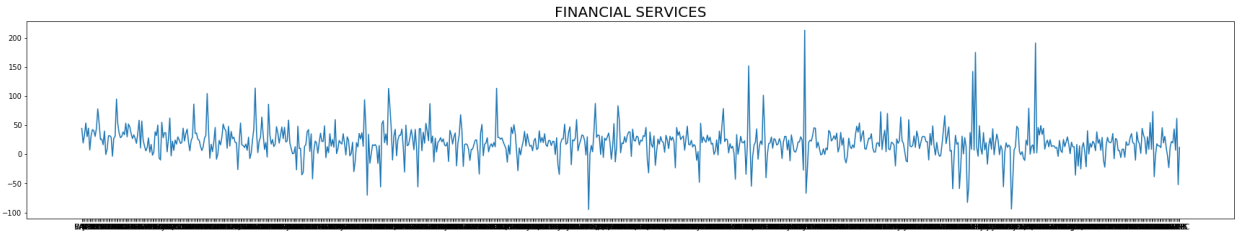


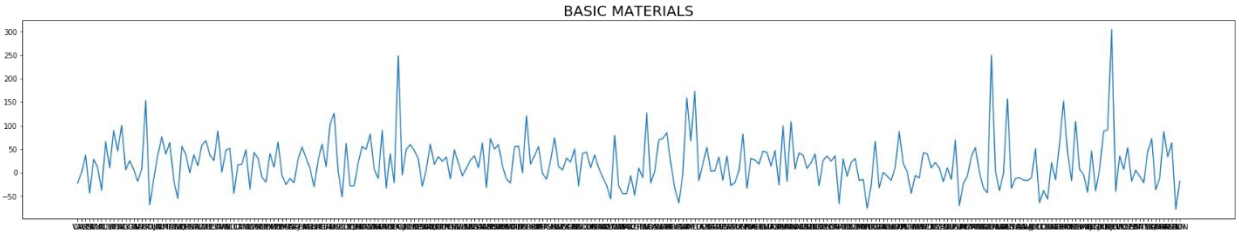


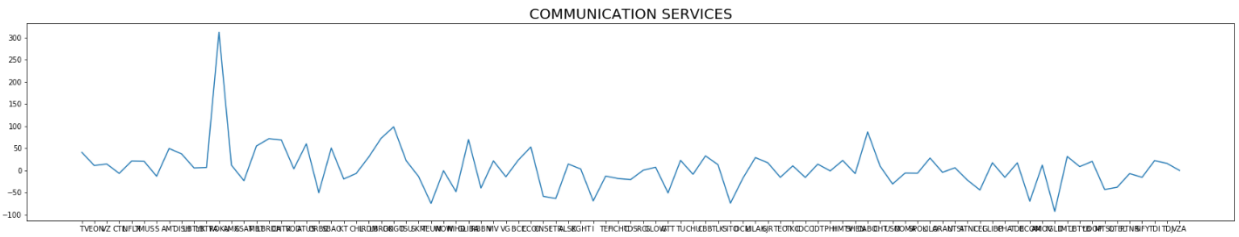
Sectors after cleaning

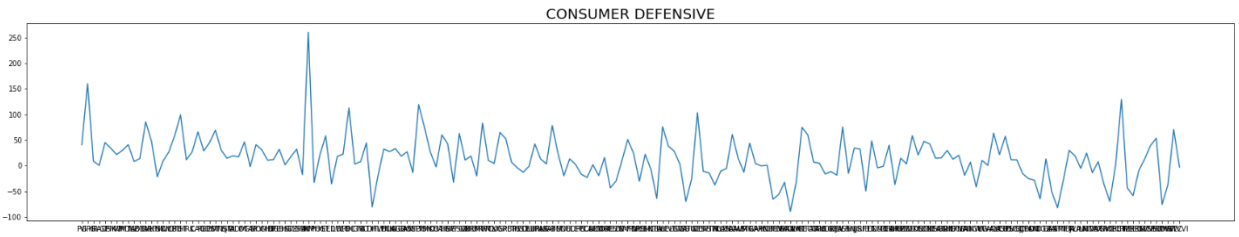


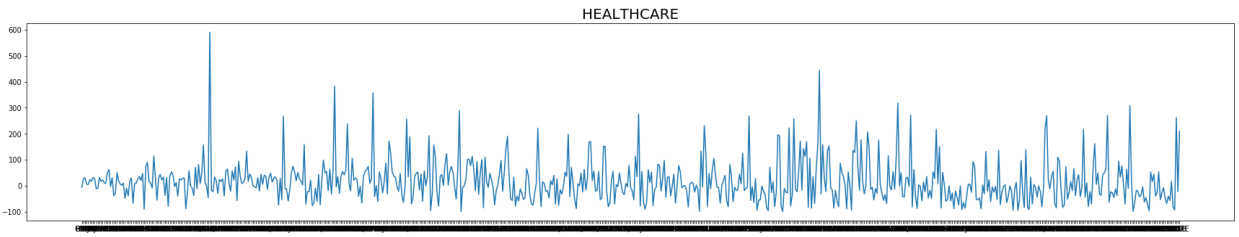

INDUSTRIALS

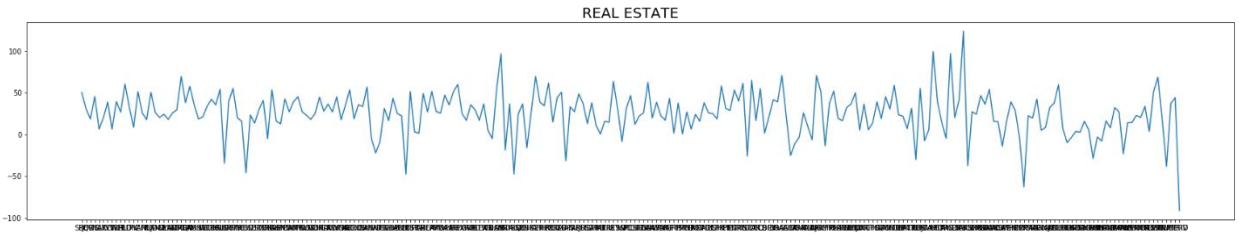

FINANCIAL SERVICES

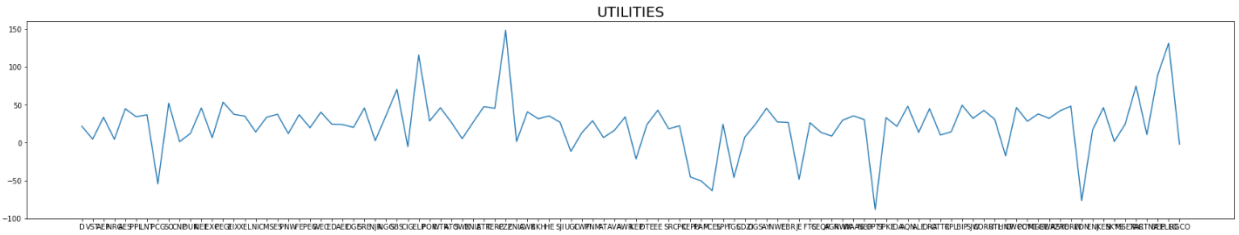

BASIC MATERIALS


COMMUNICATION SERVICES

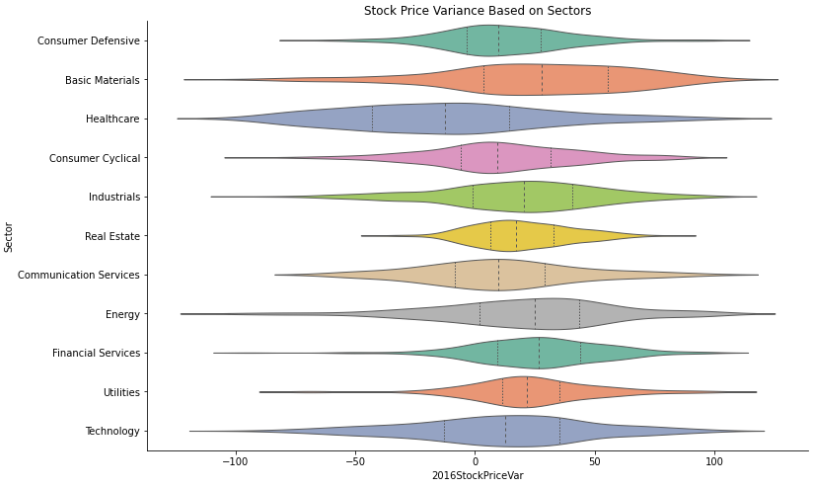

CONSUMER DEFENSIVE

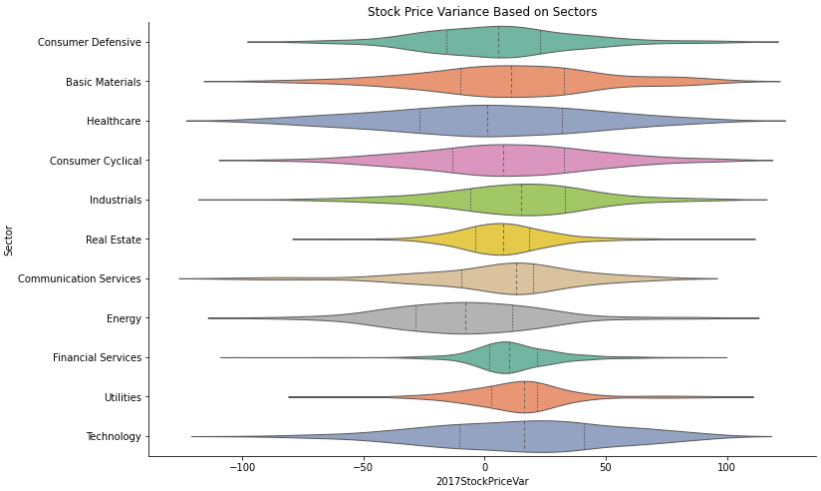

HEALTHCARE

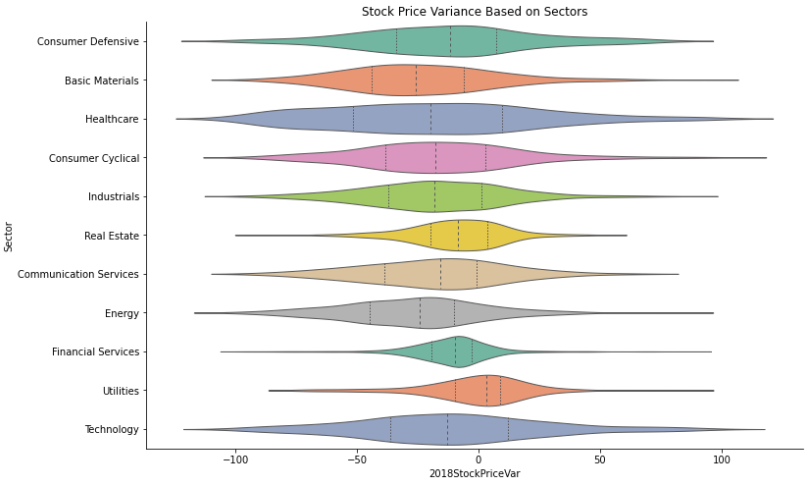

REAL ESTATE

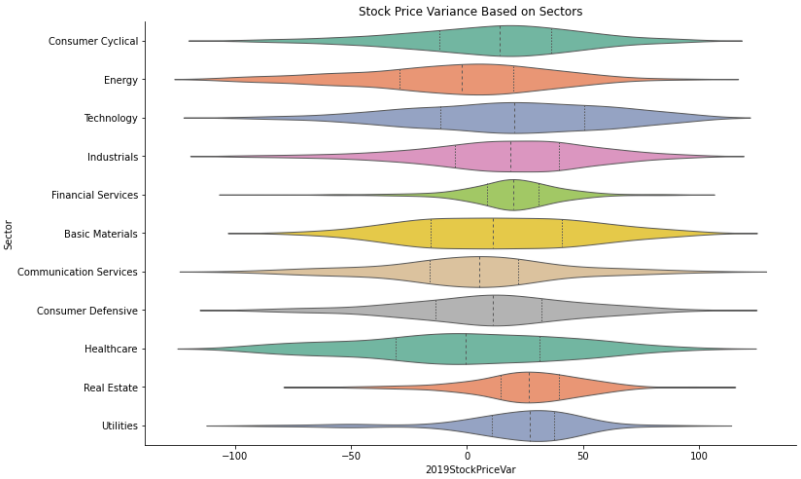


Stock price variance based on sectors








Stock Price Variance Based on Sectors
Consumer Cyclical
Energy
Technology
Industrials
Financial Services
Basic Materials
Communication Services
Consumer Defensive
Healthcare
Real Estate
Utilities
Sector
−100
−50
0
50
100
2019StockPriceVar