RANDOM FOREST AND XGBOOST IMPLEMENTATIONS TO PREDICT BANK PROFITABILITY: EVIDENCE FROM TURKISH DEPOSIT BANKS

Liva OFLAZOĞLU
Ömer Faruk RENÇBER

İKSAD
Publishing House

RANDOM FOREST AND XGBOOST IMPLEMENTATIONS TO PREDICT BANK PROFITABILITY: EVIDENCE FROM TURKISH DEPOSIT BANKS

Liva OFLAZOĞLU

Ömer Faruk RENÇBER

İKSAD Publishing House

(The Licence Number of Publicator: 2014/31220)
TURKEY TR: +90 342 606 06 75
USA: +1 631 685 0 853
E mail: iksadyayinevi@gmail.com
www.iksadyayinevi.com

It is responsibility of the author to abide by the publishing ethics rules.
Iksad Publications – 2023©
ISBN: 978-625-367-157-0
Cover Design: İbrahim KAYA
July / 2023
Ankara / Türkiye
Size = 16 x 24 cm

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Ada: Adaboost Classifier

AI: Artificial Intelligence

ANN: Artificial Neural Networks

AQ: Asset Quality

AVG: Average

BAD: Best Apparent Discovery

BC: Bagging Classifier

BS: Bank Size

CAP: Credit Assignment Path

CAR: Capital Adequacy Ratio

CART: Classification and Regression Tree

CEE: Central and Eastern Europe

CNN: Convolutional Neural Networks

CV: Cross-Validation

DNN: Deep Neural Networks

DL: Deep Learning

DT: Decision Tree

ET: Extra Trees Classifier

EU: European Union

EVA: Economic Value Added

FFNN: Feedforward Neural Network

FIAS: FinCEN Artificial Intelligence System

GBDT: Gradient-Boosted Decision Tree

GBT: Gradient-Boosted-Trees

GDP: Gross Domestic Product

GFC: Global Financial Crisis

GMM: Generalized Method of Moments

ICT: Information and Communication Technology

KNN: K-Nearest Neighbours

LR: Linear Regression

MENA: Middle East and North Africa

MER: Management Efficiency Ratio

ML: Machine Learning

MLR: Multiple Linear Regression

MSE: Mean Square Error

NB: Naïve Bayes

NBFI: Nonbank Financial Institution

NII: Non-Interest Income

NIM: Net Interest Margin

NLP: Natural Language Processing

RA: Regression Analysis

RF: Random Forest

RFE: Recursive Feature Elimination

RMSE: Root Mean Square Error

ROA: Return On Asset

ROAA: Return On Average Assets

ROAE: Return On Average Equity

ROCE: Return On Capital Employed

ROE: Return On Equity

SCP: Structure-Conduct-Performance

SEE: South-eastern European

STD: Standard Deviation

SVM: Support Vector Machine

UAE: United Arab Emirates

UK: United Kingdom

US: United State of America

VC: Voting Classifier

WWII: World War II

XGB: XGBoost Classifier

INTRODUCTION

A. Research Subject and Problem

Despite of the revived interest of policy makers in the importance of bank profitability, as a result of the global financial crisis of 2007-2009, and despite of the recovery from this crisis still many banks have their return on equity (ROE) below their cost of equity as mentioned in the Global Financial Stability Report 2016-2017.

The financial stability is surely influenced by bank's profitability; however, the literatures exhibit mixed findings when it comes to the direction of this influence. Some researchers like (Keeley, 1990; Berger and etc., 2009) have concluded that elevated profitability results in elevated "charter value" and consequently lower bank risk-taking. While others like (Natalya, et etc. 2015) suggest that high profitability may lead to higher risk-taking by loosen leverage constraints. Moreover, (Meiselman, and etc. 2018) found that the same indicator of high profit in good times, could mean contradictory a systemic tail risk in bad times. This mixed findings extend to the impact of noninterest income (NII) over risk, in studies like (Baele, De Jonghe, & Vander Vennet, 2007) or even studies like (Elsas, Hackethal, & Holzhauser, 2010) , while other like (Kohler, 2014); (DeYoung & Torna, 2013), have stated that the type of non-interest income impact the financial stability.

This shows the importance of calculating profitability and moreover, since it depends on many factors that do not affect it equally, a great value rises of evaluating the weight that each factor exerts on the profitability and, therefore, proposing a ranking of these factors

according to their importance. When it comes to classification and ranking, no better than new technologies can help us, such as machine learning (ML), a branch of artificial intelligence (AI).

AI alone have funded during the one quarter only of 2019 (the second Q), around 7.4 billion USD on a big diversity of companies and projects (CB Insights, 2019).

As defined in Britannica, AI is "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings", and it includes many specialties such as:

Machine Learning: uses algorithm to test data, which is similar to the way humans learn, and step by step improve accuracy. IBM

Deep Learning: a multi-layer of neural networks, generally three or more, that imitate the neural junctions in human brain, which result in “learning” ability from big data, the more layers there is the more the accuracy is optimised. IBM

Statistical Learning: Is a group of tools for machine learning that uses statistics and functional analysis. It aims to understand from training data to build predictive models.

Image recognition: is the ability of an algorithm to identify places, people, actions, writing, and objects in images.

Text mining: a powerful tool that can structure texts from unstructured data using natural language processing (NLP).

Social media mining: it helps extracting patterns from data generated by users on social media, allowing to take decisions about

what to advertise to users.

WEB mining: same than previous type pf mining, but instead it pulls data from world wide web to find links in between.

Internet of things: or IoT for short, describes the network of physical objects referred to as "things" that have detectors, software, and other technologies in order to be able to connect and interchange data with other devices and systems over the Internet.

In just few years ago, the financial services were relied on the physical presence of the costumer, a person-to-person service, now ML is changing the game's rules, companies and banks can decide whether to accept or reject a service or a loan demand based on historical data and algorithmic based predictions. Nowadays, the application of ML is largely infused in the market as credit scoring, automation of processes, robotic advisory services, detection of fraud, cyber security.

Nonetheless, this emergence is far from being without challenges, many risks are encountered, like bias in data, shortfalls in technology, policies and complexity. In addition to that one more thing should be kept in mind, is that the quality of data is very important to smooth running of ML's algorithms, if the data is poor, even the most luxurious system would fail.

Many literatures have been interested in bank profitability and addressed mainly bank specific factors along with macro determinants, using many classical regression methods, this study, tries to contribute to the bank profitability studies by including the most advanced ML techniques (Random Forest and XGBOOST) to obtain a ranking of

factors that are proven in literature to impact profitability, and to see how these ML algorithms perform compared with the classical Multiple Linear Regression.

B. Research Aim and Importance

This research aims at the financial level to get a ranking of the importance of the variables affecting bank profitability, and at the statistical level to compare the corelation and predicting performance of linear regression, and machine learning algorithms, as the factors affecting the bank profitability are multi-dimensional, and it is almost impossible to comprehensively study them with the conventional statistical methods, this study not only tries to correlate the variables but also ranking them using the opportunities provided by machine learning algorithms.

C. Research Method

This study focuses on Turkish deposit banks data that are five of the CAMELS rating variables (Capital Adequacy, Asset Quality, Liquidity, Activity, Size) and study their impact on bank profitability ratios (ROA, ROE), using Multiple Linear Regression, Random Forest, and XGBoost and adding a layer of Target Shuffling in the last two methods.

D. Assumptions

The period of the study extend from 2010 to 2020, and this is a very critical period as two major economic events happened, the first is the Global Economic Crisis in 2008 and the second is the Turkish debts crisis in 2018, in addition to COVID-19 in the late 2019 and beginning

of 2020, it is assumed that these factors made the environment very volatile which makes it harder for the model to set the regression and thus test their efficiency better.

E. Hypothesis

This is a comparative study, the evaluation of results will be based on R^2, MSE, the alpha level is set to 10%, the main hypothesis are:

- H_1: The models are good in predicting profitability and correlating variables.
- H_2: XGBOOST is more efficient in predicting and correlating than RF.
- H_3: A ranking can be obtained.

F. Limitations

The main limitations are the size of the studied sample with 248 observations concerning only Turkish deposit banks, and only 5 of the bank specific variables are included, other internal variables was not included like the sensitivity, and the macro determinants also.

G. Definitions

Artificial Intelligence (AI) is defined as the ability of performing cognitive functions associated with human minds by a machine, like reasoning, problem solving and even creativity (Collins. C, 2021).

Machine learning (ML) is a branch of artificial intelligence (AI) which focuses on the use of data and algorithms to learn like humans learn, an improve its accuracy (UC Berkely,2022).

Bank Profitability is the measure of a bank's performance. Banks make a profit by earning more money than what they are expending in expenses. The main source of profits are service fees, and interests from the assets (which are the loans to people and investors), while the main expense is the paid interest on liabilities (which are deposits and borrowed money from banks or commercial paper in the money market). looking at the earnings per share of the banks is not sufficient to determine the bank performance, It is also important to know how effectively and efficiently a bank is using its assets and equity for generating profits, for that three profitability ratios are to be considered while evaluating the performance of a bank: ROA, ROE and NIM (Ghebregiorgis. F, 2016).

CAMELS Rating System is an international rating system used by official authorities to rate financial institutions, based on six factors represented by the first letters, "Capital adequacy, Asset quality, Management, Earnings, Liquidity, and Sensitivity." (Kagan, 2021).

Classification is the process of grouping various entities into several classes, that can be defined according to business rules, class boundaries, or some mathematical function. Two main classes of classification can be observed according to the relationship between the elements, if a relationship is available between the reference class and characteristics of the entity to be classified, classification is called

supervised, while if no known reference classes are available, the classification is unsupervised (Nisbet. R, 2017).

CHAPTER ONE
LITERATURE REVIEW

1.1. Bank Profitability-Related Literature

Numerous studies were made to work out the factors which affect the bank's profitability:

San & Heng (2013) investigate the impact of bank-specific characteristics and macroeconomic conditions on Malaysian commercial banks performance, during the period of 2003 to 2009, The study employs pooled time-series and cross-sectional data that relate bank profitability represented by ROA, ROE and NIM. Seven variables are drawn from the conventional banking literature as proxies for bank-specific (Equity/Total Assets, Loan Loss Reserves/Total Assets, Cost to income ratio, Liquid Assets/Deposits & Short-term Funding, Total Assets of Bank) and macroeconomic factors (GDP Growth Rate, Consumer Price Index), they found that ROA is the best profitability measures. All bank-specific determinants affect bank profitability significantly in the anticipated way. However, no evidence is found in support of the macroeconomic variables have an impact on profitability.

Topak & Talu (2016) explore the factors affecting the profitability of 12 Turkish commercial banks during the period from January 2006 to March 2014, by analysic factors that can be controlled by managers, and factors that managers cannot control, the study employs Panel data analysis that relate bank profitability (ROA, ROE), to size, management efficacy, interest revenue from loans to interest expenses on deposits ratio, net fees and commisions income to total

expenses ratio,

Stockholders' Equity plus Long Term Subordinated Loans over total assets ratio, they found that only management is negativeley very significant for all the banks.

Athanasoglou et al. (2006) studied the factors determining profitability of banks in Eastern European countries like Herzegovina, Albania, Bulgaria, Bosnia, Croatia, Serbia, Montenegro, and Romania from 1998 to 2002. As dependent variables, they used ROA and ROE, and as explanatory variables, the ratio of loans to assets, loan loss provisions to loans, equity to assets, and operating expenses to assets, as well as the logarithm of assets. They found that asset profitability was negatively dependent on the ratio of loan loss provisions to loans and operating expenses to total assets and positively related to the ratio of equity to total assets and the logarithm of total assets.

Atasoy (2007) in his study, probe for the factors affecting profitability of 26 commercial banks in Türkiye, covering a period of 1990 to 2005, using bank specific, macro, and financial sector variables. They used ROA and NIM as variables of profitability. As a results of the panel regressions with bank specific variables, the ratio of equity to total assets and loan loss provisions to total assets were found to possess a positive relation with ROA but negative relation with NIM. The effect of fixed assets to total assets ratio on both ROA and NIM has been found negative. No relationship was found between deposits to total assets ratio and ROA, compared to a negative relationship with NIM.

Heffernan and Fu (2008) explore for the factors influencing the performance of banks in China between 1999 and 2006. As

performance proxies they used economic value added (EVA) with the NIM, in addition to ROAA and ROAE. Among the bank specific variables, the worth to income ratio had a negative and significant effect on proxies of performance. The ratio of equity to total assets was found to possess a positive and important impact on ROAE but was insignificant in explaining ROAA. The ratio of loan loss reserves to loans had a positive and significant impact on ROAA. The ratio of liquid assets to deposits used as a proxy of liquidity, had a negative and significant effect on both ROAA and ROAE.

Sufian and Chong (2008) in their study probe for the determinants specific for bank, also the macroeconomic of business in Philippines banks from 1990 to 2005 years. The results showed that size, risk of credit related by loan loss provisions to total loans, and overhead expenses represented by non-interest expenses had a negative impact on ROA, whereas non-interest income to total assets and equity to total assets ratios had a positive impact.

Flamini et al. (2009) explored the variables of economic bank profitability in the geographical region employing a representative group of 389 banks in 41 sub-Saharan countries over the quantity 1998-2006. They tested the impact of bank-specific and macroeconomic indicators on ROA, which they assumed to be the key profitability indicator. The ratio of loans to deposits and short-term funding used as a proxy for risk of credit had an enormous and positive impact on profitability. Capitalization was represented by the ratio of equity to total assets, it impacts positively and powerfully when used concurrently with ROA, but negatively when used with a lag of 1 year.

The rate of non-interest income to other operating revenue was used as a variable of diversification and had an opposite and highly meaningful coefficient interpreted as having a positive contribution to profitability.

Naceur and Omran (2011) investigated the determinants of bank profitability in 10 geographical region and geographic area (MENA) countries, namely Bahrain, Tunisia Jordan, Egypt, Lebanon, Kuwait, Oman, Morocco, and the UAE between the years 1989 and 2005 using the information of 173 banks. They assessed the bank performance with reference to profitability, operating performance, represented by the ratio of total operating costs to the sum of total earning assets plus deposits, and so the value of intermediation represented by NIM. Among the bank specific variables, the ratio of equity when related to total assets and credit risk defined by net loans to total assets were found to have a positive and significant effect on NIM, cost efficiency, ROA and ROE.

Taşkın (2011) sought for both the bank specific and also the macroeconomic determinants of business banks in Türkiye from 1995 to 2009. Three indicators of profitability were used ROA, ROE and NIM. As bank specific explanatory variables, the ratios of equity to total assets, total loans to total assets, loan loss provisions to total assets, staff expenses to total revenues and Napierian logarithm of total assets are used. As a result of the panel regressions disbursed, it's been found that bank specific variables are simpler in determining profitability. The ratios of loans to total assets and size have proved insignificant on both ROA and ROE. The ratio of loan loss provisions to total assets incorporates a negative and significant influence on both

proxies. Equity to total assets has been insignificant in explaining ROA, while it has a big and negative effect on ROE. The ratio of expenses by staff to overall revenues has had an opposite and significant effect on ROA, for as much as it's been insignificant in relating to ROE.

In his study, Sufian (2012) explored in addition to the bank specific variables, the macroeconomic determinants related to profitability represented by ROA on 77 Pakistani Bangladeshi, and Sri Lankan, commercial banks between the years 1997 and 2008, Based on results, the ratio of loans by total assets impact liquidity, while capitalization was impacted by the ratio of equity to total assets. Other variables were positively related to bank profitability like credit risk, the ratio of loan loss provisions to total loans, the ratio of non-interest income to total assets considered as an indicator of diversification; whereas cost expressed by the ratio of non-interest expenses to total assets had a negative and very high impact.

Capraru and Ihnatov (2014) like previous study also searched for both macroeconomic determinants and bank specific, of bank profitability over the period from 2004 till 2011, using the data of 143 commercial banks from 5 CEE countries, namely Romania, Hungary, Check Republic, and Bulgaria. They used ROAA, ROAE and NIM as dependent variables. Among the bank specific variables cost to income ratio taken as the representative of management efficiency, and equity to total assets representing capital adequacy had the strongest impact on profitability. Risk of credit were represented by the ratio of impaired loans to gross loans, was effective only on ROAA and ROAE.

Petria et al. (2015) made a study to find out the bank specific, macroeconomic and sector specific factors which determine the profitability of banks within the 27 EU countries during the amount 2004-2011. They used ROAA and ROAE as proxies of profitability. According to the empirical results, among the bank-specific variables, bank size represented by logarithm of assets, capital adequacy defined by the ratio of equity to total assets, risk of credit represented by the ratio of impaired loans to the total loans, liquidity risk represented by the ratio of loans to deposits, management efficiency represented by cost to income ratio, and business mix represented by other income to average assets ratio were found significant in explaining ROAA and ROAE.

Mamatzakis and Remoundos (2003) followed the performance of Greek commercial banks over the previous decade. They used ROA and ROE as variables and found that the deregulation of the market that happened at the time and also the procedures of European integration with the integration of the Euro have made better the competitiveness of the banking sector. The evidence clearly shows that management decisions impact the profitability of Greek commercial banks.

Athanasoglou et al. (2006) decided to add to their study the industry related determinants in addition to bank-specific and macroeconomic variables to study profitability. They used an unbalanced panel dataset of credit institutions in South-eastern European (SEE) for a period starting in 1998 and ending in 2002. The obtained results showed that, taking the liquidity as exception, all bank-specific determinants affect quit highly the bank profitability. The

macroeconomic space influences the performance of the industry. Concentration is directly correlated with bank profitability in a positive way. Inflation constitutes a strong effect on profitability, while the profits of bank don't seem to be significantly tormented by the fluctuations of GDP/capita.

Athanasoglou et al. (2008) extend the scoop to study not only the bank-specific and macroeconomic factors of bank profitability but also the industry-specific factors, the team used the standard Structure-Conduct-Performance (SCP) hypothesis to a selection of Greek banks for 16 years extending from 1985 to 2001. The results showed that as predicted, all bank-specific variables (except size) affect the profitability, however, not enough evidence was obtained to support the SCP.

Al-Tamimi (2010) in his study targeted national banks in UAE, both Islamic and traditional, during the period 1996 to 2008, the study included many explanatory variables like size, liquidity, number of branches … and GDP/capita and as dependent variable ROE and ROA. While the concentration and liquidity showed to be the most important determinants influencing the performance of conventional national banks, the number of branches and cost were shown to be the most important determinants for Islamic banks' performance.

Shaher et al. (2011) largely extended the study of explanatory factors to include 23 important factors influencing banks in the Middle East, and tried to classify them on 6 classes, for example the 1st class contained all banks characteristics, it shows the foremost important factor on profitability, this study suggests that the banks in the Mideast

region should investigate the first classes' variables to boost the performance and being able to compete with global commercial banks.

1.2. Regression and Machine Learning in Finance-Related Literature

Karabulut (2003) used the Ordinary Least Square regression along with Granger-causality tests to study the impact of the capital on bank's profitability in Türkiye, his study found that banks trying to meet capital requirements increase risky investments, and results in reducing bank profitability.

Aydın (2019) empirically and using Panel Data Analysis analyses the bank-specific, sector-specific and macroeconomic factors that affect the profitability of Turkish banks, during the years 2005-2015; It has been determined that the variables of credit risk, bank size, operating expenses, bank capital, non-interest income and economic growth are significant determinants of ROA while for ROE, variables such as bank size, operating expenses, interest income, non-interest income, inflation rate and sectoral concentration were found to be statistically significant determinants.

Yetgin and Ekşi (2017) using regression analysis bank and market-specific variables found that the size of the bank and the deposit interest rate that a bank offer have a significant effect on SME loans, while the ROA does not affect loans meaningfully.

Öz et al. (2011) used discriminant analysis to predict the stock returns for 30 company listed in Istanbul Stock Exchange for 2005 and

2006, and found that for the first year operating turnover and leverage were the more significant, while for the second year it was the operations turnover, leverage and liquidity that are the most relevent, the correct classification for second year was 91.7% and for the first year 75%.

To predict futures prices, the empirical experience of the expert is the most reliable method, yet this leads to a lot of incertitude as the knowledge of the experts' and his moods largely influence the decisions. That's why, many data mining's methods have been used to predict the price changes of financial products.

Zou et al. (2015) predicted futures prices using RF and run a comparison with guessing randomly. The three-exchange system (Dalian, Zhengzhou and Shanghai) offer the raw data to be used, each futures has nine features, 2 labels and 125 records, the period took place from July 3, 2021, to December 31, 2012. 70% of the data was randomly used for the training purpose, and the remaining 30% as the test, AVG and STD were calculated. They found that the AVG of the test varies between 59.89% and 47.35%, in contrast the random guess has only a 50% of accuracy, this study suggest that RF is capable of obtaining sound predictions with more significant results than random guess.

Xu et al. (2013) applied random forest (RF) and support vector machine (SVM) for recursive feature elimination (RFE) to select of feature based on a prediction of stock price. The stability and classification performance were investigated. The results from Shanghai Stock Exchange in China show that both support vector

machine and random forest are useful for the trend prediction.

Zheng et al. (2019) in their study used the index CSI 500 as a benchmark for all firms listed on the Chinese stock market because it represents the total performance of small-mid cap A-shares over the period extending from February 8, 2013, to August 8, 2017. The RF was used to train data, their observation suggests that ML help detecting patterns and may be of great help for quantitative trade.

Krauss et al. (2017) compared in their paper the effectiveness of three advanced ML algorithms, deep neural networks (DNN), gradient-boosted-trees (GBT), random forests (RF), in the context of predicting the one-day-forward excess return related to the S&P 500 stocks. the period of prediction was a long period of 23 years from 1992 to 2015, they promisingly found that ML can detect a profit opportunity in the short run.

Andriyashin et al. (2008) based their study upon the possibility of the relation between stocks dynamics and public information may be of nonlinear nature, thus they proposed an approach to stock choosing (XETRA DAX stocks) by employing decision trees, they found that an annual return of 25.55% can be achieved by adequately training the systems on publicly available economic data, which outperform the general market, interestingly, they found that implying variables with high volatility nature, enhance the predictability power of the model significantly.

Zhu et al. (2012) suggested that the classical parametric framework used by investors to detect relations between the performance of stock and the influencing factors is useful only if the

nature of these interactions is linear, and while the nature of these relation is of higher order and much more complexity, these parametric frameworks are less suitable. Classification and regression decision trees for stock selection were suggested, and to more explore the behaviour of these ML methods with a highly volatile set of data, they chose a data that covers the period when the pic of the Global Financial Crisis was observed in 2007 to the end of 2010 when sharp changes in investor behaviour is largely available. After comparing the classification and regression decision trees model with traditional linear framework, they concluded that the results obtained from a tree-based model was very persuasive during both the downturn in equities observed in 2007/2008 and the consequent recovery. In result and as suggested by many other studies above, the models selecting stock on the classification and regression decision trees offer more precise opportunities.

Sorensen et al. (2000) addressed in their study to answer the question What are good variables for stock selection? they stated that Traditional quantitative strategies are classes of screening techniques, although screening is helpful, however it is, by no means a complete or strictly scientific process, because some stocks may be excluded from consideration on one consideration while meeting many other considerations. thus, they introduce a replacement method to traditional ones of stock screening using statistical method known as classification and regression tree (CART). The database begins in 1992 and extends through the end of October 1999. They concluded that decision tree models lead to significantly better Sharpe ratios than do the simple

ranking models. In addition, the excess returns of all other portfolios as deducted using the simple ranking approaches prove to be statistically insignificant. In other meaning the evolving tree model performs the best.

Michal (2018) investigated many ML methods (Neural Nets, LSTMs, Q-Networks and GBT), and compared them to the Buy & Hold strategy for forecasting five different bond indexes (EUR_HY_TR, EU_Corp_TR, YS_HY_TR, US_Corp_TR and XOver_TR) from 01/01/1999 to 31/12/2017 (19 years), they concluded that, the XGBOOST and Feedforward Neural Network (FFNN) models both managed to give more reliable results than the Buy & Hold strategy on the majority of time series they were tested on. XGBOOST had on average better results, in terms of mean Sharpe ratio and standard deviation. In practice, the author recommends XGBOOST as a high yield bond index prediction model.

Elena (2021) investigates whether sentiment features resulting from financial news can enhance the classification performance using XGBOOST or not, the study used the Swedish stock market index, OMXS30 between 2006 – 2020, two sources of data were used historical prices and financial news during this period, 3,517 samples was obtained, XGBOOST was found to have a good classification performance with 73% accuracy.

Patelet et al. (2015) compare the performance of the classification of four different algorithms - RF, Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Naïve Bayes (NB) to forecast the trend of stocks and stock indices (two by two) on the Bombay Stock

Exchange, between 2003 and 2012, they found that with continuous values, RF performs with the highest classification accuracy of 83.59%, while NB performs with lowest accuracy of 73.31%. Surprisingly, when they used trend deterministic input, classification accuracy is boosted. NB reaches the highest average classification accuracy of 90.19%, and just after it, RF at 89.98% accuracy.

Ampomah et al. (2020) were interested in comparing the effectiveness of tree-based algorithms of ML models, which coincide with our thesis, they compared the results obtaining when predicting the orientation of stock price in 8 stock data from the NYSE, NASDAQ, and NSE using RF, XGB, Bagging Classifier (BC), Extra Trees Classifier (ET), AdaBoost Classifier (Ada), and Voting Classifier (VC) were randomly collected and used, as features 40 technical indicators were used, as in previous study data training 70% of the data set was attributed to the and the remaining 30% for the test, they found that XGB obtains an average accuracy of 82.66%, placing it as being the third most better performant algorithm after Voting Classifier (VC) and Extra Trees Classifier (ET), that showed the best performance, with an average performance of 83.75%.

Rongyuan (2022) in his recent study was interested in assessing risk levels of a company, XGBOOST was trained with 5 variables, bank loans, employee performance, e-commerce profit and loss, profit and loss data and cross-border business. XGBOOST model showed high reliability when it comes to assessing the risk of enterprises financials, as prediction errors were all within 3%.

Rençber And Yücekaya (2021) used ensemble methods RF, XGBoost and Bagged Polynomial Regression, to study how independent variables like Social indicators which property rights, government integrity, judicial effectiveness, tax burden, government spending, fiscal health, business freedom, labor freedom, monetary freedom, trade freedom, investment freedom, financial freedom, impact the dependant variable which is the ease of doing business index, 168 country for the period 2017-2019 gave 504 observations, the study concluded that government integrity, labor freedom and government spending are the least impacting variables.

In total, these litteratures suggest that bank profitability is mostly reflected by ROA, and ROE, with ROA being more related to profitability, the micro determinants are proved to impact the profitability in more direct way, and the four most studied variables are Management, Risk of credits, Capital adequacy, Diversification, then come Size and liquidity, but no ranking of the importance of these variables was given.

1.3. Summary of the Literature Review

In general, the studies adressed 6 banks metrics, Management effeciency expressed as operating expenses to total assets or to total revenues, Size expressed as logarithm of total assets, Capitalization expressed as equity to total assets , Liquidity expressed as liquid assets to deposits, Risk of Credits expressed as loan loss provisions to loans or to total assets, net loans to total assets, loans loss reserves to loans, loans to deposits and short-term funding, Diversification expressed as non-interest income to total assets or to other operating revenue. For

Management all obtained resluts indicate a negative impact on ROA and no effect on ROE, while for Size the results indicates eaither a negative or positive or no impact on ROA, regarding Capital Adequacy most of the studies relate a positive impact on ROA and ROE, while some found it insignificant on ROA and negative on ROE, one study adressed the Liquidity and found it negativelly related to ROA and ROE, the most adressed metric in the studies is the risk of credits, with most of the findings beeing a positive effect on ROA, lesser found a negative effect on ROA and one found it insignificant, while one found it positively and one negativelly related to ROE, finally the Diversification found to be positively related to ROA, the table 26 in the Conclusion and Discussion section summerizes the findings of the literatures and this study regarding the direction of the correlation.

When it comes to ML application, most of the usage of advanced techniques is for stock price prediction and highly volatile environment, some studies included RF, other included XGBoost along with other methodes, only Krauss et al. (2017) included both of RF and XGBoost but also in stock price prediction, the importance of these two methods are that they are the most advanced techniques used in ML. This study included tries to expand the scope from the previous studies and use both RF, and XGBoost as being the most advanced ML models to suggest a ranking of the main micro-determinants factors influencing the profitability.

CHAPTER TWO
THEORETICAL REVIEW

2.1 General Finance System

2.1.1 Financial Market

It is an organised structure that mediate buying and selling of financial assets or other services, including loans, stocks, bonds, deposits, options and futures, it provides the needed credits of the individuals, firms and institutions, in other words it facilitates the transmission of funds between investors (savers) and the borrowers. The financial markets can be classified into money market and capital market, the money market is a market for short term securities like call money, treasury, bills, bills of exchange etc., raised to satisfy the short-term requirements, offer a safe investment of money but with low rate of return, it plays a role in maintaining balance between demand and supply of short term funds, and making funds available to many units with diversified activities, also it helps the growth of economy by providing funds to developing sectors, and it provides a base for the implementations of policies, the capital market is a market for long term securities such as equity, debentures, bonds etc. (Cecchetti & Schoenholtz, 2015).

2.1.2 Financial Institutions

These are structures that mobilize savings and provide finance or credit to individuals and organizations as described above, it can be classified into Banks, Non-banking Institutions, and Specialized institutions. The Bank’s main role is to take deposits from those who have money and give them as loans to investors who need

funds, in this role banks are intermediaries between depositors and borrowers, they profit form the difference between amount banks pay to depositors for their deposits and the income they receive from investors, this profit is called interest and determine the banks main profits source. Nevertheless, the nonbank financial institution (NBFI) as the name indicates, it is not full banks and thus does not have a full banking license, which means that the public cannot deposits their money in, however, NBFIs play a role in facilitating other financial services, such as risk pooling, investment (both collective and individual), brokering, financial consulting, check cashing, and money transmission. In addition, they are a source of consumer credit. As Examples insurance firms, some microloan organizations, currency exchanges, fit in this category of financial institution, as someone can conclude, the services these institutions can deliver are not necessarily exclusive for banks, thus competes with banks, and specialize in sectors or groups. Specialized institutions provide medium- and long-term credit to industrial investors, they are multipurpose institutions which discover investment projects, provide technical advice and managerial services and assist in the management of industrial units, like banks for agriculture and agricultural cooperatives, government housing banks etc. (Cecchetti & Schoenholtz, 2015).

2.1.3. Financial Instruments

These are financial assets comprising instruments from money market and capital market, the first are short term financial assets like call money, notice money, treasury bills, certificate of deposits, commercial papers etc., the second are long term financial assets

consisting of equity shares, preference shares, debentures, bonds etc. (Cecchetti & Schoenholtz, 2015).

2.1.4. Regulatory Agencies and Central Banks

The responsibility of regulatory agencies is to ensure the proper functionality of the financial system – including its institution, market, and instruments, while the central bank's main role is to the monitoring and stabilisation of the economy (Cecchetti & Schoenholtz, 2015).

2.1.5. The Principles of Money and Banking

Time Has Value: time affects the value of financial transactions, a loan when spread over time will end up with much more than the loan value itself, the difference is a compensation of the borrower for the time.

Risk Requires Compensation: some of the financial risks can be eliminated, some can only be reduced and mitigated, but some needs to be taken care by another institution, which will be compensated by explicit money, this is the example of the car insurance for example which will shoulder the risk someone doesn't want to take.

Information is the Basis for Decisions: the foundation of the financial system is the collection and processing of information, for example before releasing a loan, the officer will investigate the financial status of the costumer, to make sure he/she is able to pay back.

Markets Determine Prices and Allocate Resources: the core of the economic system are markets, it is the meeting point for buyers and sellers, either physically or virtually, the financial market is a crucial

corner of the economy, the more the financial market of a country is developed the faster the country will grow, and this is due to the fact that markets determine prices and allocate resources.

Stability Improves Welfare: as mentioned before, the main role of the central banks is to stabilise the economy by controlling inflation and reducing business cycle fluctuations, in other words, keep the inflation low and stable, and keep the growth high and stable (Cecchetti & Schoenholtz, 2015).

2.2. Regression Analysis

Regression Analysis (RA) is a set of statistical processes working as modelling system that is useful mainly to evaluate the relationship between dependant (Y or response or target) and independent (X or regressor) variables, and for predicting values of dependant variable using non-existing (in the sample) values of independent variables. It is the Sir Francis Galton, who was the first to use the term “regression”, nevertheless Adrien-Marie Legendre and Carl Gauss was the first to developpe the first form of RA called method of least squares (Bingham & Fry, 2010).

The regression line, is the best fit line that best describes the relations of a dataset, it is called the best fit because it minimizes the residuals between the line (estimated values) and the point (real values), being a line allows to deduct its function easily, with single slope (b) for linear regression and multiple slopes (b1, b2, …, bn) for multiple linear regression, and intercept (a0) regression lines are mainly used for forecasting, they are widely used in finance to predict prices of

stocks, commodities and perform valuations, moreover regression lines are used in businesses sector for forecasting sales, inventories, business strategy and planning (P.Sarkar, 2022)

$Y = b\,X + a_0$ for single linear regression.

$Y = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n + a_0$ for multiple linear regression.

The difference (y-$\hat{y}$) between a dot on this line, which is the estimated value ($\hat{y}$) and the true value of the dot (y) is the error, or residual, the less it is, the best the model is, and as it can be negative or positive, it is squared to eliminate the negative sign, and to magnify the errors, the sum of these squared error is SSE which must be as low as possible.

$$SSE = \sum_{n}^{i=0} (y_i - \hat{y}_i)^2$$

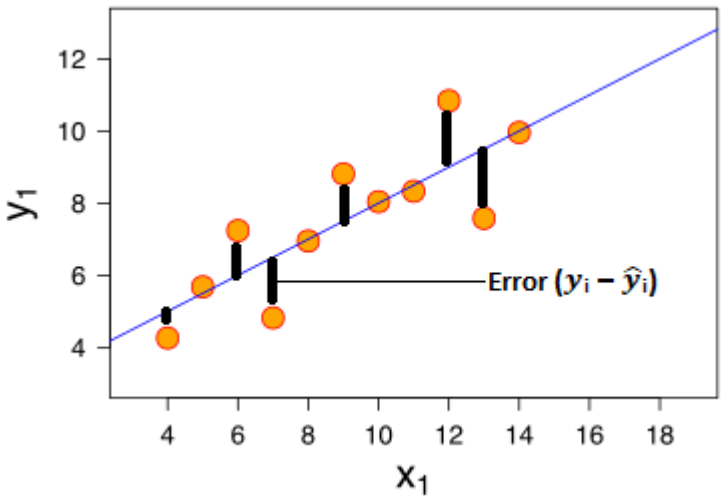


Figure 1. *Residuals or Errors (medium.com).*

The Ordinary Least Square (OLS) method for linear regression is the mathematical method used to estimate the best line slope and intercept that minimize the SSE to the best.

The bivariate normal distribution is indicative of normal distribution, the regular Normal Distribution is about one variable that randomly changes, while the bivariate consider two variables (x,y) changing randomly, If x and y have a standard bivariate normal distribution, y depend on x linearly, and has the equation of the regression line.

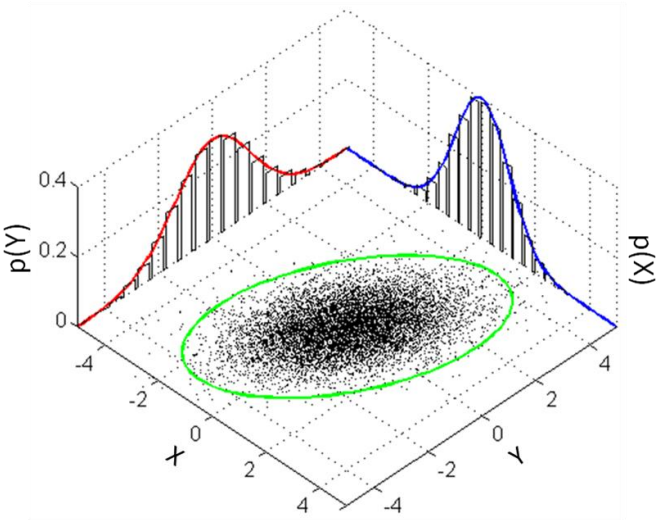


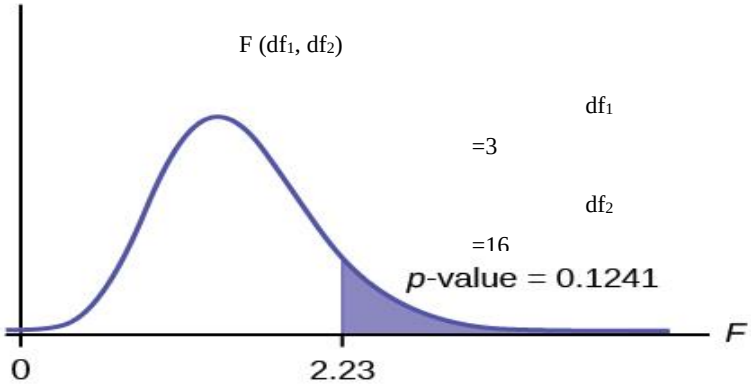
Figure 2. *Multivariate normal distribution (wikipedia.org).*

The ANOVA test is a statistical test that allows to compare more than two groups simultaneously to determine if a relationship exists between them and to what extent, the one-way ANOVA is used when there is one independent variable, while in the case of two independent variables it is called two-way ANOVA, and when the study contains

several dependant variables it is called MANOVA, the results of the ANOVA test are reliable if only some assumptions were met (R. G. Miller, 1997).

- The population distribution needs to be close to Normal Distribution.
- Samples needs to be independent.
- Variables need to be homoscedastic (equal variances).
- Groups with equal size.

The F-test is the main output of The ANOVA test, it allows to reject the null hypothesis (H_0), that is there is no relationship between variables, or not to reject it, the F has two degrees of freedom (df_1, df_2) that allow to calculate the F critical value (corresponding to α=0.05) that the F value needs to be bigger in order to reject H_0, sometimes it is possible to calculate the significance F (the probability of H_0 being true) that correspond to the F value, and compare it with α=0.05, if the significance F (*p*-value) is smaller than α=0.05 H_0 is rejected.



The *p*-value when F > 2.23 (Critical F) = 0.1241

Figure 3. *Example for P-value is bigger than α, the H_0 cannot be rejected* (Illowsky & Dean., 2012).

2.3. Artificial Intelligence and Machine Learning

In only fifty years, AI have offered significant applications, AI is capable of doing things that humans do but do them intelligently.

2.3.1. History and Evolution of AI and ML

During history three big industrial revolutions can be observed. The first started in 1784 with the invention of engines moved by steam. The second was in 1870 when electricity began. While the third revolution in 1969 correspond to the IT revolution. Now the fourth industrial revolution is being witnessed, AI which is about big data (Skilton & Hovsepian, 2017).

Table 1. *Industrial revolution.*

Industry	Year	Description
Industry 4.0	Today	Networks, Cyber Physical Systems, internet of things,
Industry 3.0	1969s	Automation, electronics, and computers
Industry 2.0	1870s	Massive production, electrical energy, assembly line
Industry 1.0	1784s	Mechanization, steam, power, weaving loom

The third revolution has offered necessary technology to burst the fourth one with a mixture of new technologies that melt the limits between computers, physical, and biological environments (Schwab, 2017).

In the beginning, between 1950s and 1960s the majority of the AI

work was oriented to Bayesian statistics, although the Bayesian statistics has prepared the field for the ML being used today, it has at the time no applications in the financial services, the biggest hinder was lack of technology for data and storage. Soon the AI in 1970 entered the "AI Winter". After only 10 years and with better computer technology and new funding's opportunities, AI witnessed a revival in the 1980s. The first AI application in financial industry was forty years ago, in 1982 when James Simons started his firm "Renaissance Technologies". Sooner In the 1990s, AI showed big importance within fraud detection. Systems were implemented to detect money laundering like FIAS, which can scan 200 000 transactions weekly, the results were astonishing, 400 laundry's attempts worth $1billion was detected within two years (Senator, et al., 1995).

Back again to the main hindering to the AI at the time, computer processing power and storage, in 2011 with new processing and storage technologies, new possibilities were open especially within DL which has become the breakthrough of AI. Although the main growth of AI has taken place in the US with 71.78% of the world's total funding (Buchanan & Cao, 2018), but the US has lost their placing when China started to invest heavily and overrun the US in overall funding (CB Insights, 2018).

The figure shows how Global hubs outside USA is taking over the AI deals between 2014 and 2018 to switch from only 25% of total deals to nearly 60% in just 5 years, which reflect the arising worldwide interest in AI products.

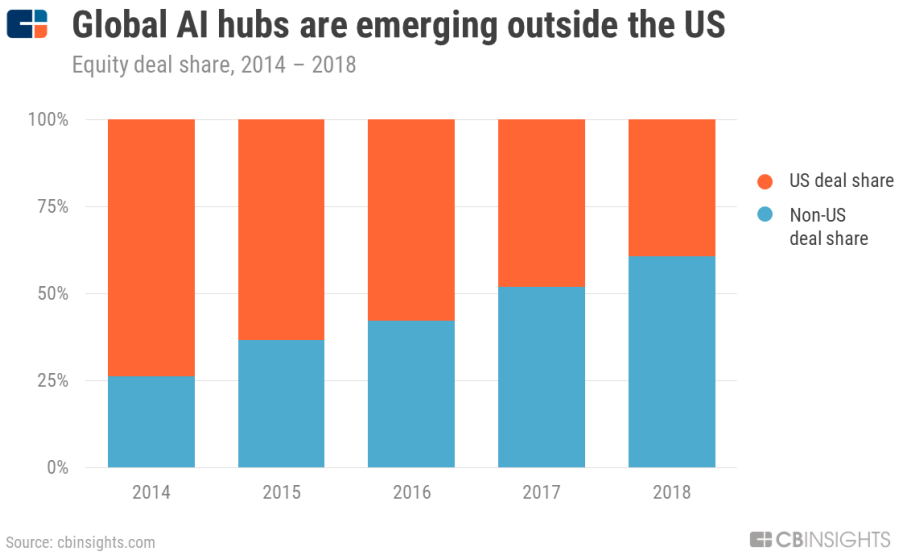


Figure 4. *Global AI hubs outside the US & cross-border AI deals (cbinsights.com).*

It is useful to mention that the amount of data plays a big role in the advances of AI, today around 2.5 quintillion (1018) bytes of data are manipulated every day, which is 2.5 time the metres wide of Milky Way Galaxy, 90% of these has been generated during the last two years alone (Marr, 2018).

Finally, Europe is behind in this race, the ICT sector stands for only 1.7% of the GDP with contrast to 2.1% in China, and only 1.65% of the US GDP, knowing that Europe's GDP is close to the US and is little bit higher than China’s, this rise the question about why it is so, especially that there are around six million developers in Europe which is more that the developers number in the US, European Commission has put up a fund collected from public and private investments to

invest around €20 billion each year from over the next decade (European Commission, 2019) .

2.3.2. Types of AI

Giving a short explanation about the four types of AI will help as mentioned before to see where ML fit and understand the opportunity and the future of these technics. With reference to Forbes, four types of AI are present: reactive machines, limited memory machines, theory of mind, and self-aware AI (Joshi, 2019).

Table 2. *Types of Artificial Intelligence.*

1. Reactive machines
2. Limited memory
3. Theory of mind
4. Self-aware

Reactive machines as the name suggest, these machines can only respond automatically to a limited input, they do not have the ability to learn, the IBM's Deep Blue is an example of these machine, which by responding to the movement of pawns on a chess board won over Garry Kasporov the chess Grandmaster in 1997. The Google AlphaGo is also another example of this type (Ray, 2018).

Limited memory machines contrary to the reactive machines which don't have functionality based on memory, can retain data but only for a short time, this allows them to learn from archived data and to make future decisions. Nearly all known applications nowadays fit in this category, ML and DL algorithms rely on historical data to generate

an output, so all the AI application these days are derived from limited memory machines, The most famous example is “Sophia” made by Hanson Robotics in 2016. the humanlike robot that can understand ideas, emotions and react accordingly, with a social behavior like joking or asking reciprocal questions, these algorithms can run a full conversation with human with all its feelings, intentions, expectations, and motives components (Reynoso, 2019).

Lastly, the extension of theory of mind is the Self-aware machines, this type of AI doesn’t exist yet, but it is considered to be the most advanced form of AI known to humans. In contrast with all previous AI technics that react to inputs only, this machine think for its own, expresses passions and comprehend what feelings are, these are the type of dangerous robots seen in the movies, they are super intelligent and conscious that some researchers say that these intelligent machines may be a serious threat and form a high risks to us humans (Yaninen, 2017).

2.3.3. Overview of Artificial Intelligence

AI ameliorate the computers processing enabling them to act like humans. In spite of the absence of machines that are clone of humans, but AI has advanced enormously in recent years. the ultimate is to develop smart machines that are autonomous, in interpreting data and learning. In order to create accurate forecasts far better than the human based forecast, a machine needs to analyse high dimensional data, and detect patterns, this is what AI does easily, AI can be divided to 6 main categories as shown in table 3 (Mardanghom & Sandal, 2019).

Table 3. *Overview of AI categories.*

1. Machine Learning
2. Neural Networks
3. Robotics
4. Expert Systems
5. Fuzzy Logic
6. Natural Language Processing

2.3.3.1. Machine Learning

This is maybe the foremost cornerstone of AI. ML enables systems to learn and develop automatically from experience without the need to be explicitly programmed. It depends on t accessing data and analyse it to find out patterns and learn (Expert Systems, 2017). The learning process or training looks like computational statistics and starts with information's observations. Then a pattern is detected, tested, and then used to predict.

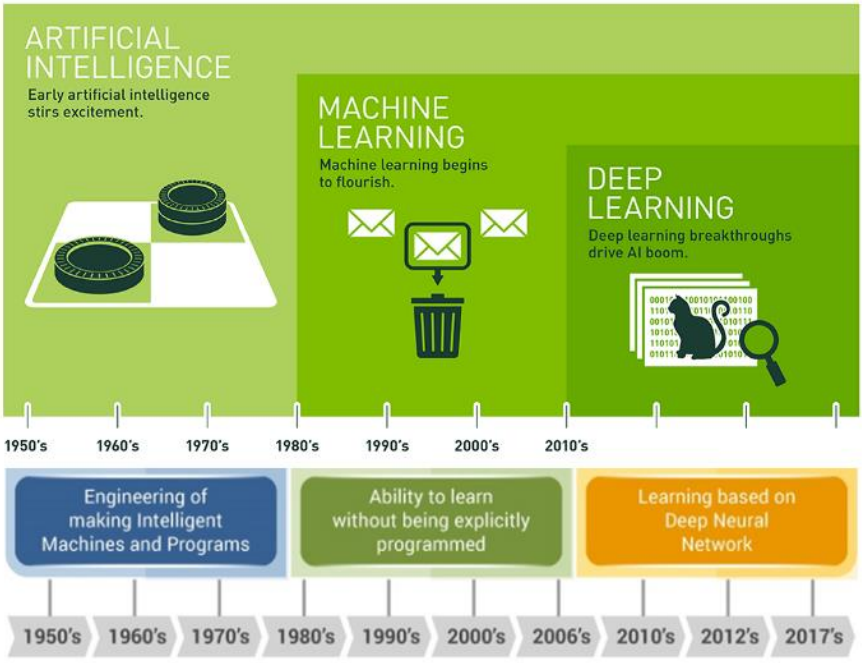


Figure 5. *History of AI (ai.plainenglish.io).*

The ML's algorithms are divided to four types, supervised ML, semi-supervised ML, unsupervised ML, and active ML. ***Supervised ML*** for example when the machine is needed to detect if a set of picture contains a bridge, or financially speaking which image contains an Iban number for example. ***Unsupervised ML*** algorithms in the other hand, are used if data is not classified. For example, when a machine is requested to detect a loan and payback patterns for a bank customer or investor. ***Semi-supervised ML*** as the name indicates, it uses both supervised and unsupervised. Lastly, ***Reinforcement ML*** the machine is rewarded with a signal if it reveals a controversy in many ways (Dietrich, Heller, & Yang, 2015).

Table 4. *ML classification* (Dietrich, Heller, & Yang, 2015).

Machine Learning				
Supervised Learning		Unsupervised Learning		Reinforcement Learning
Regression	Classification	Clustering	Dimensionality Reduction	
Advertising Popularity Prediction Weather Forecasting Market Forecasting Estimating Life Expectancy Population Growth Prediction	Diagnostics Customer Retention Image Classification Identity Fraud Detection	Customer Segmentation Targeted Marketing Recommender Systems	Feature Elicitation Structure Discovery Meaningful Compression Big Data Visualisation	Game AI Skill Acquisition Learning Tasks Robot Navigation Real-time decisions

2.3.3.2. Deep Learning

The algorithms of ML are considered as not deep or "shallow" because the input can easily surpass some layers. In image processing for example, lower layers can identify edges, in comparison higher levels can reveal letters, humans like digits, or faces (Techopedia). In DL, computers try and do what humans do naturally. An algorithm is considered deep if a non-linearity is detected before an output is given (Schmidhuber J. , Deep Learning, 2015).

Table 5. *Comparing an ML approach to a DL approach* (Schmidhuber J. , Deep Learning, 2015).

Machine Learning	Deep Learning
Input	Input
Feature Extraction	Feature Extraction and Classification
Classification	
Output	Output

The term DL refers to the number of layers through which the information is processed, each layer contributes on the results of previous layers to the next ones, for example the first layer may detect tires, the second street, the third a shape of car and the third conclude that this photo is a car on the street.

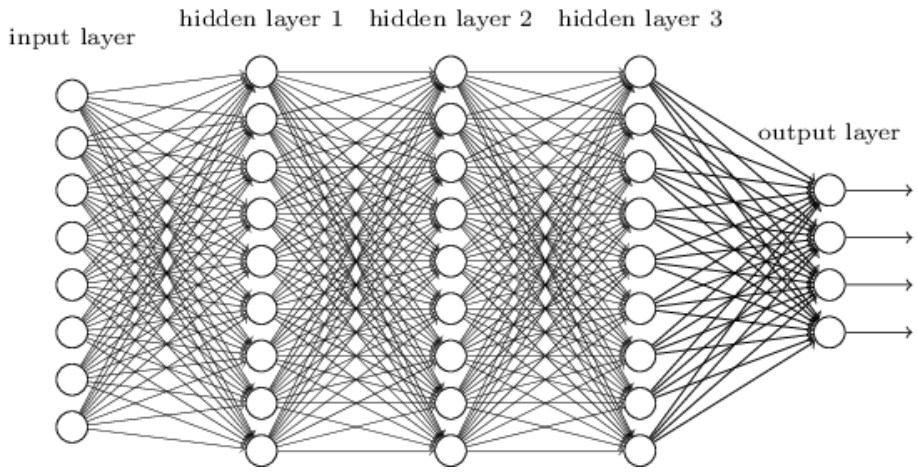


Figure 6. *Neural Networks and DL (neuralnetworksanddeeplearning.com).*

2.3.3.3. Big Data

When data sets are divers and ever growing, the term big data is used, these big data can't be analysed by the commonly used software do to its size and diversity, they are used to predict user behaviours, when massive data is brought up, four main attributes need to be remembered, (Korpela, 2017), to whom two more attributes was added recently.

How much generated and stored data is the Volume. How many types of information is the Variety, as Big Data is extracted out from divers sources like text, audio, pictures, and video. Velocity indicates the speed of generation and processing of information. Veracity, the quality, and the degree of certitude that the information gives. Variability concerns the degree and speed of the growing of the structure and data.

Recently two more attributes were added to the previously cited, the Value which indicates the business importance of the collected data, and Variability which means how data can be used and formatted. It is obvious that the structure of AI and the abilities make it so convenient for big data, that they are now seemingly inseparable.

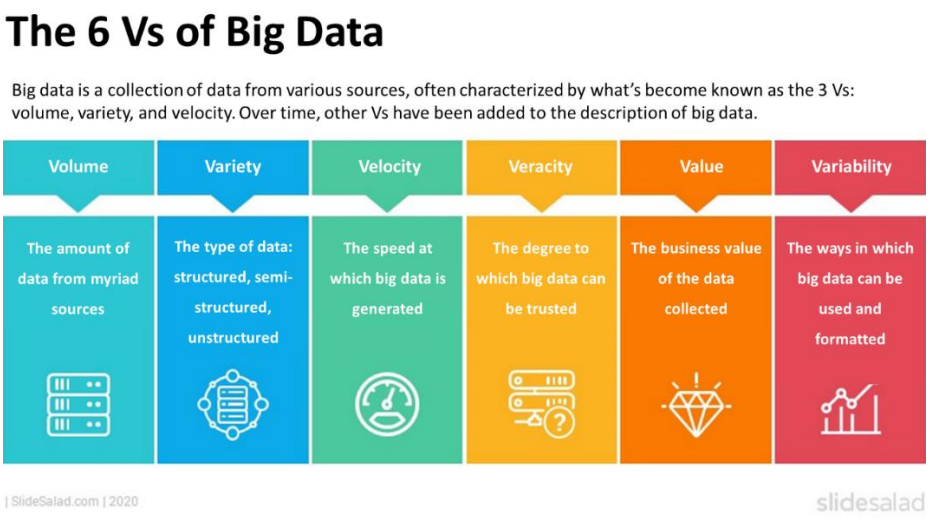


Figure 7. *6V's of big data (slidesalad.com, 2020).*

2.3.3.4. Random Forest Regression

One of the biggest roles of ML is to classify tasks, this has a big value in business applications. For example, using previous data to classify and decide whether the loan customer will default or not. In ML, there are different classification algorithms like k-nearest neighbours (KNN), Naive Bayes, Logistics Regression, and Decision tree, but the RF classifier is considered the best when it comes to classification tasks.

RF is a supervised algorithm, although it is the easiest and most used algorithm for classification tasks, it is also used for regression tasks, it is preferable to go through the concept of decision trees first as it is the fundamental that is used in the Random Forest classifier.

Decisions Tree is a powerful method to take decisions in complex scenarios. The method consists of using certain parameters to

repeatedly divide information into different categories.

In DT, the procedure's flow follows a tree structure according to conditions. The structure consists of internal nodes, branches, and a terminal node. Internal nodes test an attribute, while the branches hold the conclusion of the test, and terminal node or leaf node, means the class label. Sometimes the name "CART" is used as it means Classification and Regression Tree, they are always preferred due to the stability and the reliability.

An example of a very basic decision tree will be set, where a decision of whether to play cricket or not, as said before, a decision tree is composed of:

- Root Node - Represent the whole sample that is going to be divided.
- Decision Node - gets further divided into different sub-nodes.
- Branches/Sub-tree – all Divisions are called branches.
- Parent and Child Node - Parent node is the original node; child node is the node derived from a parent node.

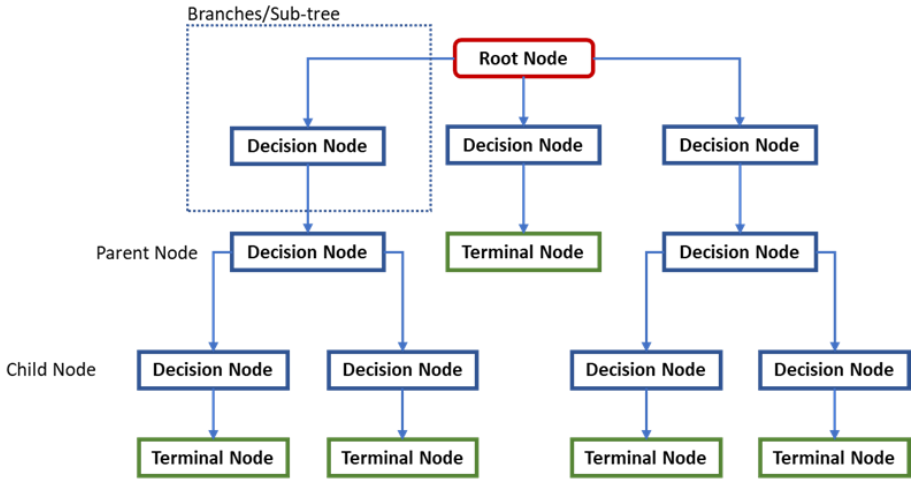


Figure 8. *Decision tree components.*

It is the answer yes or no that decide which branches to go with, the algorithm starts with the root node and then makes a comparison of the value of different attributes and continues the next branch until it reaches the end leaf node (Terminal Node). Different algorithms are used to check the homogeneity of the split and variable, In the Cricket example a decision tree would look like this:

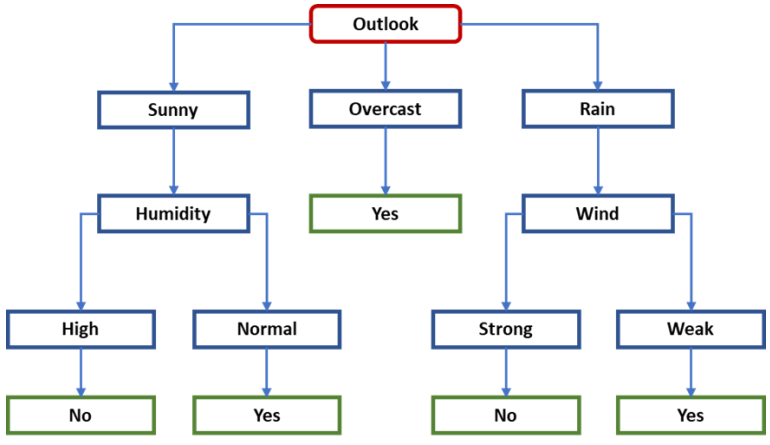


Figure 9. *Decision Tree for playing cricket.*

Depending on the type of data a categorical or numerical decision tree can be obtained, in the previous example if the decision is only yes/no, it would be categorical, or also called classification tree, but if the humidity or wind for example are evaluated by numbers, and the decision is based on the value of these numbers, it would be a numerical, or continuous variable tree, or also regression tree.

Advantages

- Very simple and effective.
- Applicable even if there is missing values in the dataset.
- Suitable for numeric as well as categorical features.
- Direct results, no need for statistical or mathematics to be explained.

Disadvantages

- Small changes in training data can result in transformation of the logic vent.
- Difficulty to interpret larger trees.
- Bias can be occurred if having more levels.

For Random Forest, when an algorithm generates different decision trees and then collect them randomly in groups of votes, it is called a random forest, a forest because the decisions are based on many trees, and random because the generation of these trees is random. Indicators like gain, gain ratio, and GINI index help to form the decision tree, then the average of all the outputs from all trees is calculated and presented as a result, the steps are:

- Choose samples randomly from the dataset.

- Generate DT related to each sample and calculate prediction results from each decision tree.
- Calculate votes for each predicted result.
- The final result (prediction) is the result that has the maximum votes.

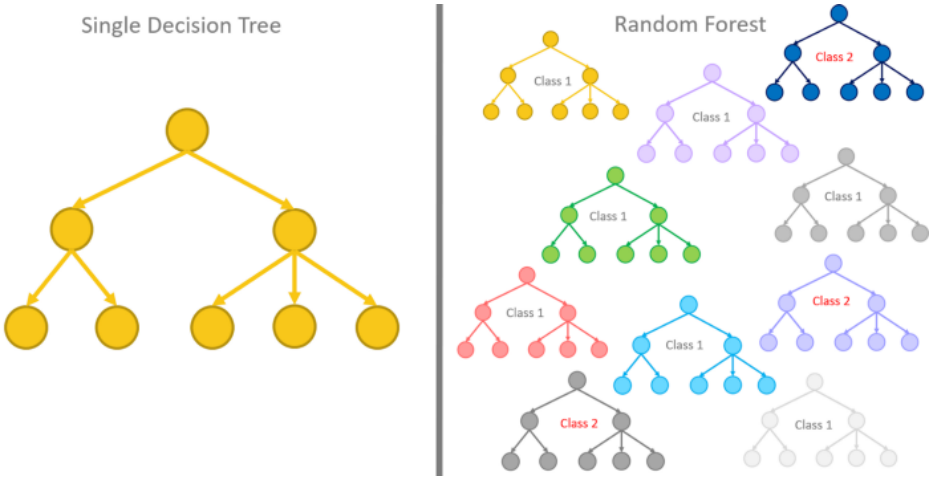


Figure 10. *Decision Tree, and Random Forest* (Silipo, 2019).

Advantages

- It can be used for both classifications (categorical) as well as regression (numerical).
- No overfitting if there are enough trees.
- Fast and can handle missing values.
- It is flexible with high accuracy.

Disadvantages:

- not easy to interpret due its complexity.
- Time consuming compared to other algorithms.
- High computational resources are needed.

2.3.3.5 XGBOOST

Ensemble learning is the core of XGBOOST, this is a way for assembling the abilities of many learners to predict, in a systematic way. in other words, the final outcome assembles the result of many models in one model. The boosting is provided parallel, the term gradient boosting refers to the fact that one weak model is assembled with many other weak models to create a stronger model, this minimises errors from one model to the other.

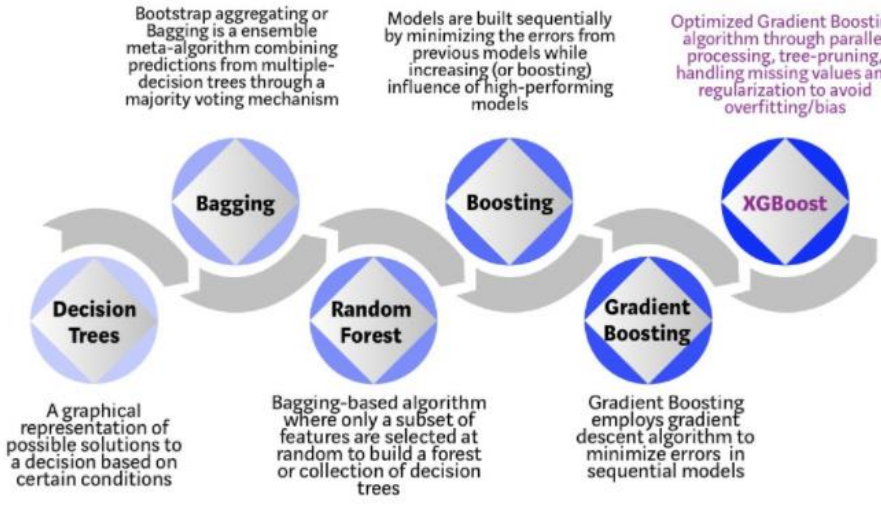


Figure 11. *Evolution of XGBOOST (Morde, 2021).*

Whether it is for classification or regression problems, XGBOOST is the most largely used algorithm in machine learning. Its good performance as compared to all other machine learning algorithms, the reasons behind this good performance are:

- A dominant factor of the algorithm is the regularization used

to get rid of the overfitting.

- Missing values can be managed.
- Flexibility that gives support to objective functions, which are used to evaluate the performance.
- Saving and loading again, which saves time.

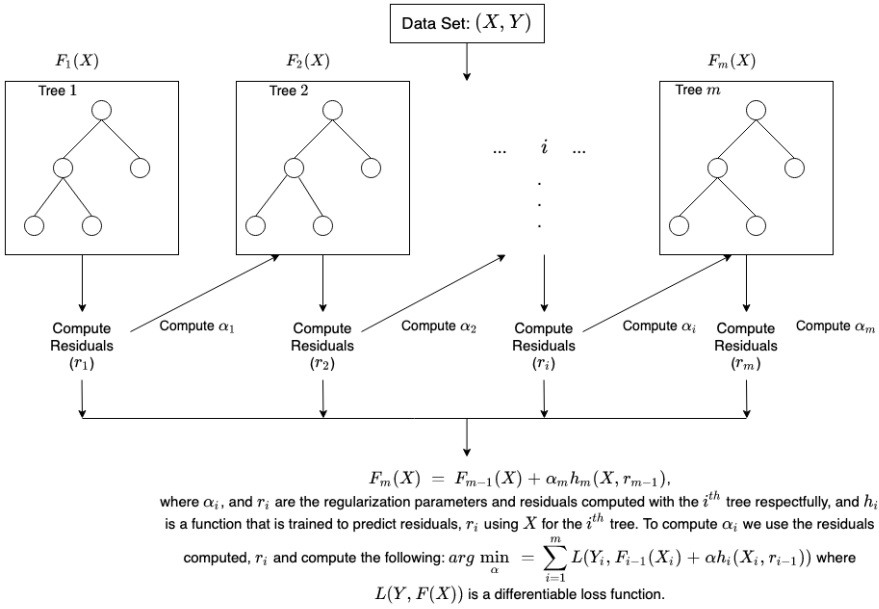


Figure 12. *XGBOOST working (Amazon 2021).*

XGBOOST data pre-processing steps:

- Information loading.
- Eliminate any unnecessary attributes.
- Converting text to numeric values.
- Identify and replace missing values.
- Splitting dataset into training and testing.
- Scale the features or normalise the data.

- Do Principal component analysis process, which is a statistical process that transform correlated variables to uncorrelated variables, using orthogonal transformation.

What makes XGBOOST more efficient than gradient boosting?

The first answer is the used regulation, which is more regularized in XGBOOST compared with Gradient Boosting, using advanced regularization (L1 & L2), allow boosting the generalization capabilities of the model. This allow the deliverables of XGBOOST to be of high performance as compared to Gradient Boosting. The second answer is that the XGBOOST training is greatly fast and can be executed parallel over clusters. The third answer finally, is that the missing values are handled internally.

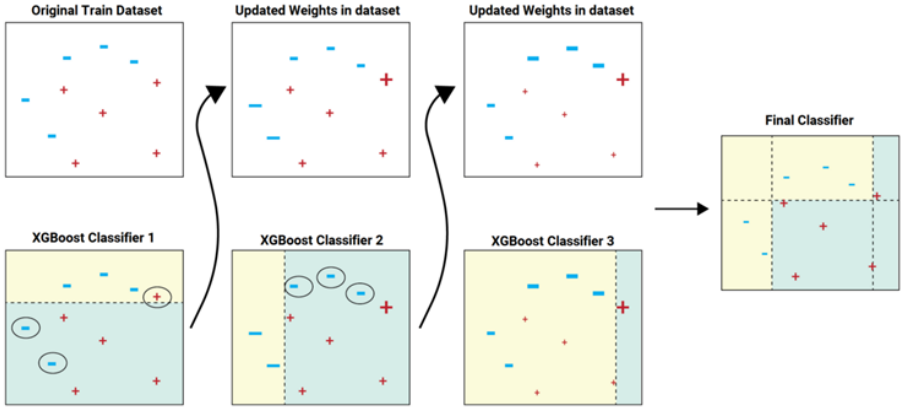


Figure 13. *XGBOOST classifier (AlmaBetter,2021).*

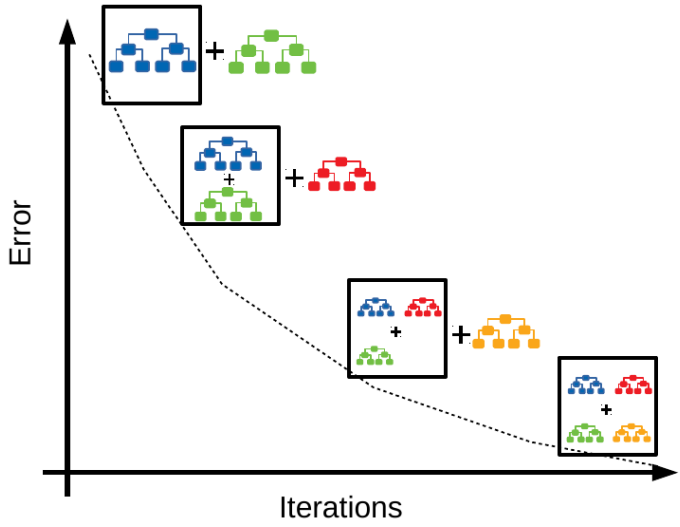


Figure 14. *Gradient boosting adds sub-models incrementally to minimize a loss function (tvas.me, 2019).*

One last thing to note is the tree pruning way that the XGBOOST uses, 'max_depth' rather than criterion, and pruning is started backward, it is a technique of data compression to minimise the size of decision trees by eliminating unnecessary sections of the tree.

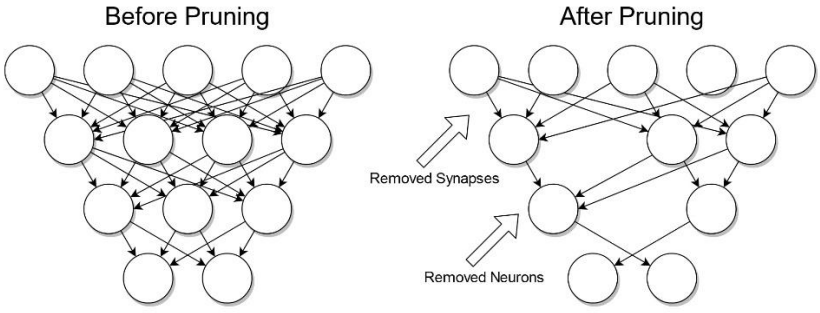


Figure 15. *Tree pruning (Wikipedia, 2014).*

In conclusion, both RF and XGBoost uses decision trees, but the way data flow throw the algorithm is different, RF generates different decision trees and then collect them randomly in groups of “votes”, so the model is built upon votes, while XGBoost after each tree, it compute the “residuals” and feeds it to the second tree, so that the errors diminish from a tree to another, this what help the algorithm to perform faster and more in a direct way. This study tries to see if this difference in performance between the process of random “votes” in RF, and computing “residuals” in XGBoost will lead to a difference in the correlation and the prediction for the data set in this study, and which is the best performing model.

CHAPTER THREE
VARIABLES DEFINITIONS

3.1. Return on Assets (ROA)

Net incomes found on statement devised by total assets found on balance sheet, collectively admitted as the most relevant profitability measure within banking sector. It expresses the ability of a corporation to generate revenues from its assets and thus profits, one bank may have higher net income, yet if it has a low ROA, it is less efficient at making profits than another bank with higher ROA, even if this last bank has much less net profit, that is why ROA allows investors, economists to make more clear assessment of banks and take decision about investments. A ROA over 5% is considered as good, and above 20% is considered as excellent. This metric is of great help to compare the profitability of different institutions within the similar industry, this condition is met in this study as all studied institutions are banks working in the same sector, however in case of institutions working in different fields, the measure may not be as precise, most of the studies about profitability uses ROA as mentioned in our theoretical review (Hargrave, 2022).

$$ROA = \frac{Annual\ Net\ Income}{Total\ Asset} \ldots (\text{Roman}, 2013)$$

3.2. Return on Equity (ROE)

Profit over Total Equity, in other hand reflect the profit drawn from the invested capital by shareholders, it is quite important for investors as they need to judge how efficiently the institution is able to use their investments to generate additional revenues, a steady increase of ROE over time means that management is able to generate positive value for shareholders, many experienced investors look for a ROE equal or more than 15%. However, it cannot be used alone as a separate metric, because it doesn't take in consideration financial debt or leverage (Prep Waal Street) (The Economic Times).

$$ROE = \frac{Annual\ Net\ Income}{Average\ Shareholders'\ Equity} \ldots (\text{Roman}, 2013)$$

As from the balance sheet, the asset equals to the liability plus the equity, the equity could be calculated as:

$$Equity = Asset - Liabilities$$

If the bank had no liabilities at all, the equity would equal to the assets, ROA and ROE would be equal as the denominators are equals. However, if the liability gets bigger, the equity (ROE denominator) diminishes and consequently ROE would rise, and as the asset (ROA denominator) rises by taking debts, the ROA will diminish, this shows well how liability can amplify ROE and shrink ROA. Both ROA and ROE provide a clearer picture of how bank's management is effective in generating profits, If ROA is high enough and the debt level is reasonable a good ROE will indicate that the managers are doing good

job in generating profits from shareholder's investment (Mcclure, 2021).

3.3. Capital Adequacy Ratio (CAR)

The proportion of Capital (Tier 1 or core capital + Tier 2 or the second layer capital) to Risk weighted Asset; it's considered as a percentage reflecting the bank's risk weighted credit exposures, that is why it is critical to make sure that bank has enough cushioning capacity to stand before reasonable amount of loses before reaching a degree when banks lose depositors' funds, it is also used by regulators to run stress tests, the upper this ratio is, the safer the bank is and the less is the need for shareholder's equity. However, there is one limitation that it fails to reflect the bank's expected losses during its run (Hayes, 2022).

$$Capital\ Adequacy\ Ratio = \frac{Tier\ 1\ Capital + Tier\ 2\ Capital}{Risk\ Weighted\ Assets} \ldots \text{(Roman, 2013)}$$

3.4. Asset Quality (AQ)

AQ is the total loans and receivables over total assets, loans are considered as assets for banks, the interests earned from these assets are the main income, while the main risk is the risk of some of these loans not being paid back, a loan with high credit risk is a low quality loan, or low asset quality, because bank needs to hold more capital to weight the risk of credits, in any economic crisis, the asset quality ratio is key metric to observe, because many borrowers fail to payback their loans. (Europian Central Bank). Asset quality reflects risk that banks could

face, the more non-performing loans are included in bank's portfolio, the more the exposures to failures are likely (Gunsel, 2007).

$$Asset\ Quality = \frac{Total\ Loans\ and\ Receivable}{Total\ Asset} \ldots \text{(Roman, 2013)}$$

3.5. Liquidity

Although Liquidity has many types, current ration, quick ratio, cash ratio, basic defence ratio and basic liquidity ratio, this study include the liquidity as the Liquid assets over Total Assets, liquidity is a very important part for any institution as it is required for it to pay off its short-term obligations, it measure the efficiency of the institution in converting its assets to meet urgent needs, and helps understanding how cash is easy to be available, the higher this ratio is the more efficient is the institution to clear debts, it is a very important ratio especially for creditors check before the obtaining of short term loans. The higher the liquidity than 1 the better the liquid position is, however, this ratio has some limitations to be taken in consideration, first it considers only the current assets, that is why it is better to also consider other metrics, second there is a risk of overestimation of inventory needed to calculate the liquidity, third it may be the result of creative accounting because only the balance sheet's information is included (Tuovila, 2021)

$$Liquidity = \frac{Liquid\ Asset}{Total\ Assets} \ldots \text{(Roman, 2013)}$$

3.6. Management Efficiency Ratio (MER)

It is the operating expenses over the total asset; it reflects the

capability of an establishment to well utilize assets and liabilities internally, this ratio can impact the institution's activities in many ways:

- The Allocation of projects and strategies is a very important managerial activity to create value from resources, a failed allocation dramatically reduces the management efficiency by spending capital to activities that generates no value.
- Productivity is the one hour working employee output.
- The efficiency in using resources such as energy, land, water … etc. without wasting.
- The efficiency of using time, labor and money in a particular process, like for example putting ATM machines inside banks to allow customers drop and deposit more that the limit of the externa ATM machines can allow.
- The efficiency of cost, what a unit is costing to generate the desired output.

As it is shown above, the MER is a very important metric to estimate how well the institution is using its resources to generate profits. (Spacey, 2018)

$$MER = \frac{Operating\ Expenses}{Total\ Asset} \dots (\text{Roman}, 2013)$$

3.7. Bank Size

Total Asset is represented by BS in most articles or thesis. Generally, is expressed by the natural logarithm of the total asset. Usually, it impacts profitability positively, which means that the

larger the size, the more the profits (Athanasoglou, Brissimis, & Delis, 2008).

$$\text{Size} = \ln(\text{Total asset})$$

Nevertheless, Size itself behave differently with risks, it is independent variable when comparing large (over 50 billion USD assets) and small banks, larger banks are more exposed to risks than smaller one, while BS ceases to be independent risk factor when comparing large banks only, in this case the insufficient capital is the main risk driver (Laeven, 2014).

CHAPTER FOUR
APPLICATIONS

4.1. Aim and Scope

Three main aims of this study are:

1. **Statistical**: compare the performance of two of state-of-the-art of ML, RF and XGBOOST in predicting and correlating variables, with classifying financial data.
2. **Financial**: obtain a rank that shows the importance of each independent bank-specific variable that influence profitability's main indicators; ROA and ROE, mainly Capital Adequacy, Asset Quality, Management Efficiency, Liquidity, and Bank Size.
3. **Macroeconomic**: beside the fact that there is a significant relationship between banks profitability and GDP (Reis, 2016), this study address if the application of ML can be helpful in monitoring and observing the effect of ROA and ROE on Macroeconomic.

For this, three main hypotheses were formed:

1. H_1: The models are good in predicting profitability and correlating variables.
2. H_2: XGBOOST is more efficient in predicting and correlating than RF.
3. H_3: A ranking can be obtained.

These hypotheses are based on the assumptions:

1. ML algorithms are more efficient in studying nonlinear

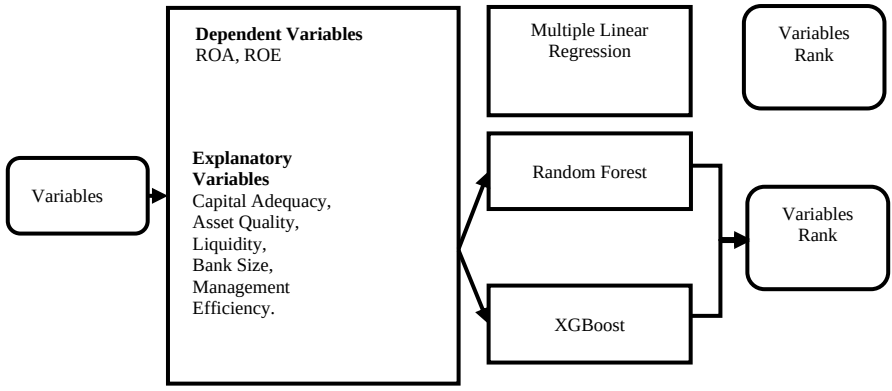
regression

2. where many independent factors can be taken in consideration to study simultaneously their effects on the dependant variables.
3. XGBOOST is more efficient than RF as it uses gradient boost to eliminate false model, in this way less and less errors are included in the model, thus the name GRADIENT.
4. The negative impact of COVID-19 on the small to medium businesses which are the customers of deposit banks, may impacted their ability to pay back debt, which may raise the risk and lower the profitability.

4.1.1. Importance of The Study

Many studies have evaluated the factors influencing bank profitability, using a method like statistical regression, panel regression and other, this study aims to rank the importance of those influencing variables over profitability using advanced machine learning methods. The theoretical frame is presented in table 6.

Table 6. *Theoretical frame.*



The period of the study extends from 2010 to 2020, during this period Türkiye has been under two economic depressions, the Global Financial Crisis in 2008, and the Turkish Debt Crisis in 2018, these crises made the economic environment very volatile and instable and surely impacted the banks profitability, the GDP annual growth percent went up and down from a maximum of 11% in 2011 to a minimum of 1% in 2019.

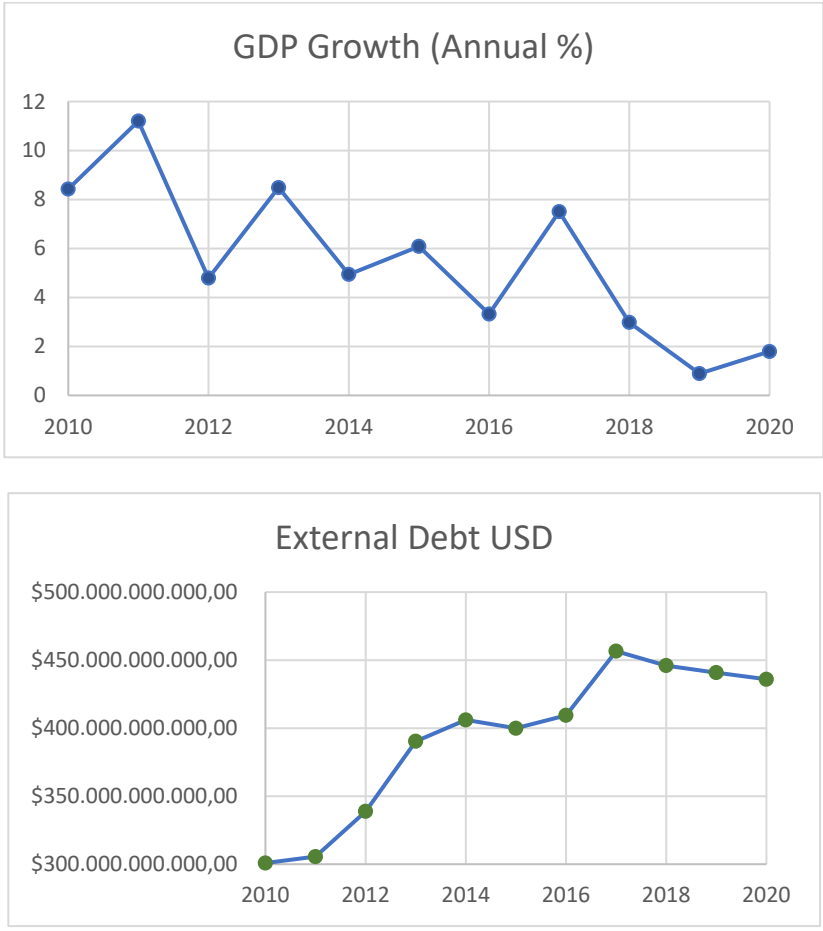


Figure 16. *Turkish GDP and External Debts during the study period (worldbank.org) (macrotrends.net).*

The inflation rate was sort of stable from 2010 to 2016 with values not surpassing 8%, but in 2017 jumped to 11% and reached a higher pick in 2018 to be 16%, this inflation came along with a steady change in the Turkish Lira value compared to US Dollars, to go from 0.5 TL/USD in 2010 to 7 TL/USD in 2020.

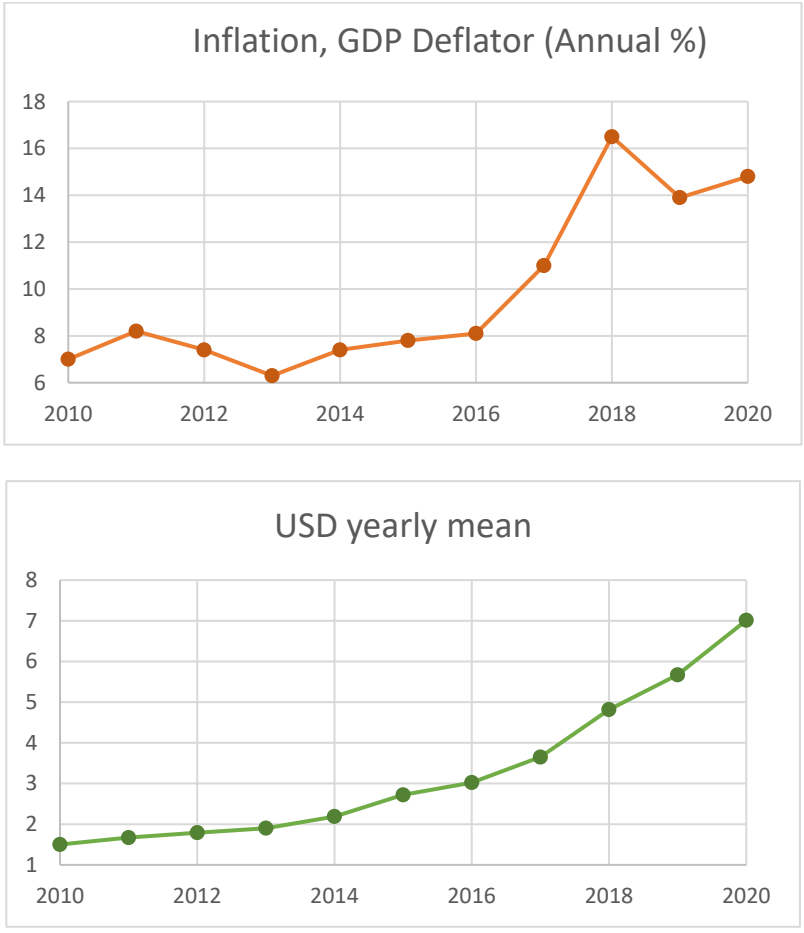


Figure 17. *Turkish inflation rate and USD exchange rates during the study period (worldbank.org) (exchangerates.org.uk).*

In addition, we cannot neglect the impact of COVID-19 pandemic that was declared officially by WHO on 30 January 2020 as public health emergency of international concern and a pandemic on 11 March 2020. In fact, this highly volatile economic and health environment is in favour of the study, as more volatile data are more difficult to be analysed and the harder is to correlate variables, and the

better are the models tested.

4.2. Dataset and Method

4.2.1. Dataset

The dataset was obtained from The Bank Association of Türkiye (The Union of Turkish Banks), Report Code: YE05 issued in July 2021 and covering the amount 2010-2020.

The data contained ratios for 48 Banks as mentioned in, the bank system in Türkiye consists of two main categories of banks, Deposit Banks, Development and Investment Banks. The deposit banks are divided into State-owned Banks (3 banks), Privately-owned Banks (8 banks), and Banks Under Depo. Insurance Fund (2 banks), Foreign Banks (Foreign Bank Founded in Türkiye 16 banks, Foreign Banks Having Branches in Türkiye 5 banks). The Development and Investment Banks are divided into State-owned Banks (3 banks), Privately owned Banks (7 banks), and Foreign Banks (4 banks).

This study concerns only Deposit Banks; from which only State-owned Banks (3 banks), Privately owned Banks (8 banks), and Foreign Bank Founded in Türkiye (16 banks) in a total of 27 banks. In this study Banks Under Depo. Insurance Fund and Foreign Banks Having Branches in Türkiye which belong to Deposit Banks too were excluded because the bank risks are influenced, and only banks that have a ratio for the years 2010 to 2020 were included, thus Adabank A.Ş. data were excluded as only Assets Quality Ratios were available, other ratios weren't. The final list of banks included in this study is shown in Table 7.

Table 7. *Banks included in this study.*

State-owned Banks	**Foreign Bank Founded in Türkiye**
Türkiye Cumhuriyeti Ziraat Bankası A.Ş.	Alternatifbank A.Ş.
Türkiye Halk Bankası A.Ş.	Arap Türk Bankası A.Ş.
Türkiye Vakıflar Bankası T.A.O.	Bank of China Türkiye A.Ş.
	Burgan Bank A.Ş.
Privately-owned Banks	Citibank A.Ş.
Akbank T.A.Ş.	Denizbank A.Ş.
Anadolubank A.Ş.	Deutsche Bank A.Ş.
Fibabanka A.Ş.	HSBC Bank A.Ş.
Şekerbank T.A.Ş.	ICBC Türkiye Bank A.Ş.
Turkish Bank A.Ş.	ING Bank A.Ş.
Türk Ekonomi Bankası A.Ş.	MUFG Bank Türkiye A.Ş.
Türkiye İş Bankası A.Ş.	Odea Bank A.Ş.
Yapı ve Kredi Bankası A.Ş.	QNB Finansbank A.Ş.
	Rabobank A.Ş.
	Turkland Bank A.Ş.
	Türkiye Garanti Bankası A.Ş.

4.2.2. Method

For the Multiple linear regression, Data Analysis by Excel was used, the main attention was drawn to R squared, F test and its significance, and the *P*-value for each independent variables, along with the *t*-statistic which will help to rank the variables. While for algorithmic regression, KNIME was used as an ML platform, it is an intuitive open-source software permanently assimilating new developments for generating data science, attention was drawn to R square and MSE before and after target shuffling. After organizing data in the proper form to be ready for ML purposes, the information was fed to KNIME for 8 Modules, 4 before Target Shuffling, and 4 after.

4.2.2.1. R Squared:

R-squared is how good is the measure of correlation between dependant and independent variables in the regression models. This

parameter indicates the percentage of the variance in the dependent variable that can be explained by the independent variables, in other words R-squared determines how strong is the relationship between the model and the dependent variable on a convenient 0 – 100% scale, the higher R-squared is the better the model is.

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

R^2 = R-squared

SSR = sun of squared of residuals.

SST = total sum of squares.

4.2.2.2. *F* Test and Its Significance:

When it comes to multiple linear regression, the model suggests a linear equation that represent the best fit of all independent variables together, which will allow to predict further values, the multiple linear regression is of the form:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n + a_0 + u$$

Y: dependant variable.

X: independent variables

b: respective slop of every independent variable.

a_0: the interception with Y axis.

u: the residuals of regression

the F-statistic is used to test the linearity of variables, in other

world if at least one of the independent variables exhibits linearity, and that means that this variable explains a significant portion of the variation of the dependent variable. The F test is useful then to calculate the *P*-value (significance of F test) which allows to accept or reject the null hypothesis, which states that there is no linearity, or all variables show a slope ($b_1=b_2=b_3=\ldots=b_n=0$) of 0, if *P*-value is smaller than 0.05. then the null hypothesis is rejected, and the existence of linearity is proved, otherwise the null hypothesis cannot be rejected, and the regression model couldn't be valid.

$$F = \frac{MSR}{MSE}$$

MSR: Mean sum of squares.

MSE: Mean squared errors.

$$MSR = \frac{RSS}{K}$$

RSS: Regression sum of squares.

K: Number of independent variables.

$$MSE = \frac{SSE}{n - K - 1}$$

SSE: Sum of squared errors.

n: Number of observations.

Both k and n-k-1 are called degree of freedom of F.

The t-value measures the importance of the independent variables, by reflecting how much the dependant variable can be affected by this specific independent variable, the greater the value of T, the greater the influence of the variable and the greater the evidence against the null hypothesis, the sign of t indicate the direction of the influence.

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

$\overline{x}$: The sample's mean.

μ_0: Th hypothesized population mean.

s: The sample's standard deviation.

n: The sample's size.

4.2.2.3. Before Target Shuffling:

Module one, and Module two concerning Random Forrest for ROA and ROE respectively, whereas Module three and Module four concerned XGBOOST for ROA and ROE respectively. Each module consists of 5 nodes, Reader, Partitioning, Learner, Predictor, Numeric Scorer, the scale of the first partition was set to 70%, so as to permit 30% for testing, at the tip of the Mean Squared Error (MSE) from the Numeric Scorer's statistics was registered for each of the four modules to be used as reference.

4.2.2.4. After Target Shuffling:

The same four Modules were added one additional node; Target Shuffling after the Reader and before the partitioning, then it had been

set to focus on explanatory variables one by one, Capital, Asset, Liquidity, Activity, Size, for every variable MSE was recorded. The difference between the first calculated MSE taken as a reference with the obtained MSE for every variable is calculated then, the greater this difference the more important is that the variable.

4.2.2.5. Mean Squared Error

MSE reflects how close or distant a regression curve is to a group of points; in other words, it reflects how good is the model in predicting. It takes the distances from the points to the regression curve (these distances are the "errors") and square them. The squaring is critical as it transform negative signs to positive, and also amplifies the differences. It's called the "mean" as it reflects the common of a collection of errors. The lower the MSE, the higher the forecast.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

MSE = mean squared error

n = number of data points

Y = observed values

$\hat{Y}$ = Predicted values

4.2.2.6. Root Mean Squared Error

Although the MSE is a metric to judge which is the best model, the RMSE is straightforward to interpret so it's used more often, it is the root of the MSE, though it has the same unit as the variable, that's why it is easier to interpret, In practice, several regression models are

applied to a dataset and RMSE is calculated of each model, then the "best" model is selected based on the lowest RMSE value, because it is the one that makes the best predictions that are closest to the actual values from the dataset, in this study the dependant variables are themselves a percentage, the RMSE is reflected in % though.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$$

4.2.2.7. Target Shuffling

According to (Elder Reasearch) for data science, AI and machine learning, to test the statistical accuracy of data mining results the Target Shuffling process can be used, It is especially useful for identifying type one error (false positives), or when two events happening together are regarded to have a cause-and-effect relationship, as opposed to a coincidental one. The more variables there is, the easier it becomes to 'over search' and identify (false) patterns among them this is called the 'vast search effect', the process consist of:

1. Building a prediction model for the target variable and evaluate the model strength.
2. break the relationship between output and its vector of inputs by randomly shuffling the target vector.
3. Looking for a new best model or best apparent discovery (BAD) and save its strength.
4. Repeat steps 2 and 3 many times and record a distribution of the strength of the BAD's.
5. Note where the finding's strength lies on this BAD

distribution, that is its significance.

6. For example, if 15% of the BAD's are better then there is a 15% chance that the original finding is bogus.

The calculated MSE before shuffling reflects well the prediction taking in consideration the relations between outputs and inputs, while after shuffling, this relation is broken, and the calculated MSE should be bigger as more bigger residuals are expected to show up, though the difference between the two MSE, before and after shuffling will has two possible meanings:

1. If the difference tends toward Zero, then the initial model before shuffling is not or very little different from the model after shuffling, this conclude to an initial model with broken relationships between inputs (independent variables) and outputs (dependant variables), and any detected pattern or relation is highly due to coincidence.
2. The bigger the difference, the higher the difference between the initial model without shuffling and the model after shuffling, which means that the relations between inputs and outputs detected in the initial model are highly due to a real cause-and-effect relationship, otherwise it couldn't be so different from the shuffled model.

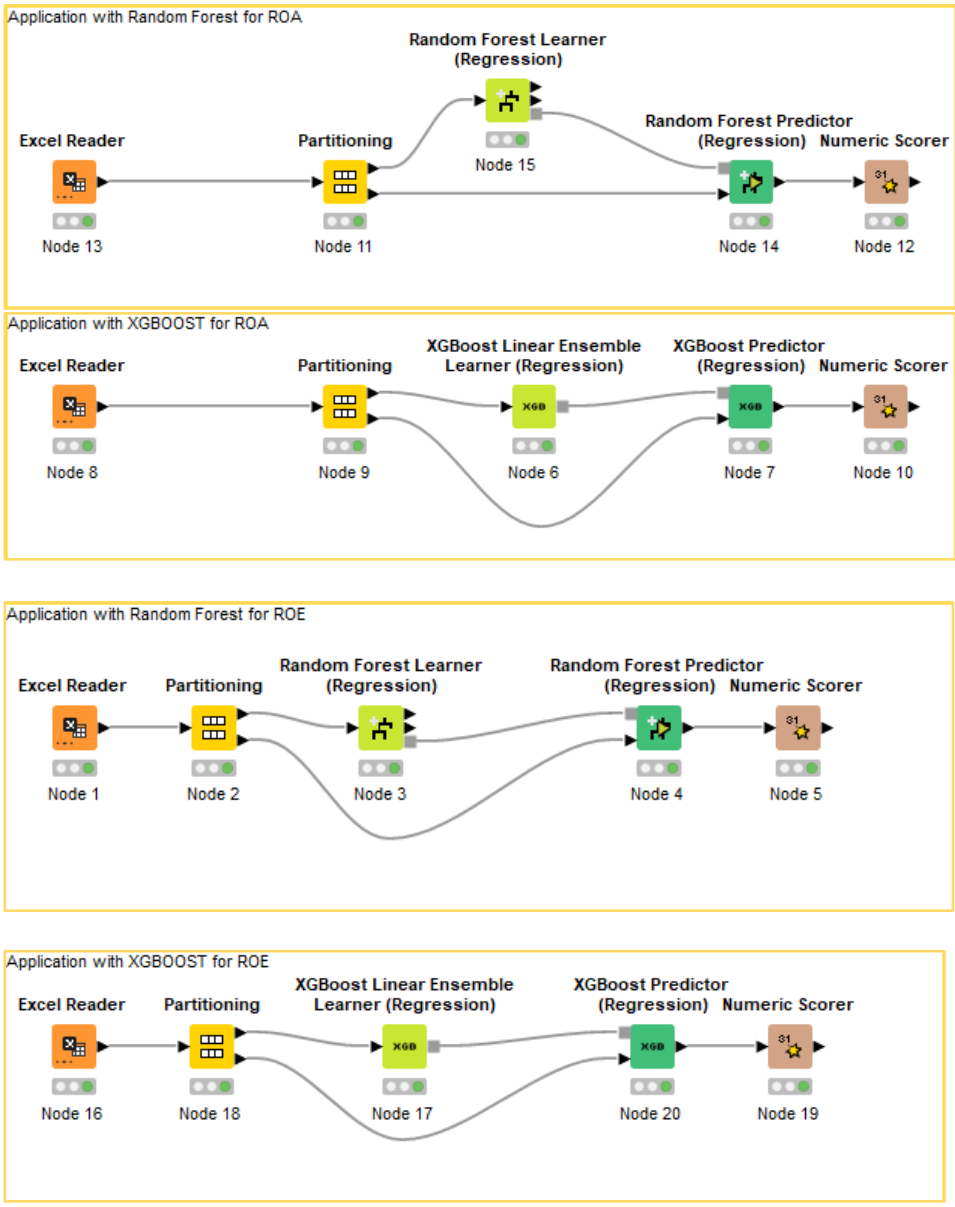


Figure 18. *The four models before Target Shuffling.*

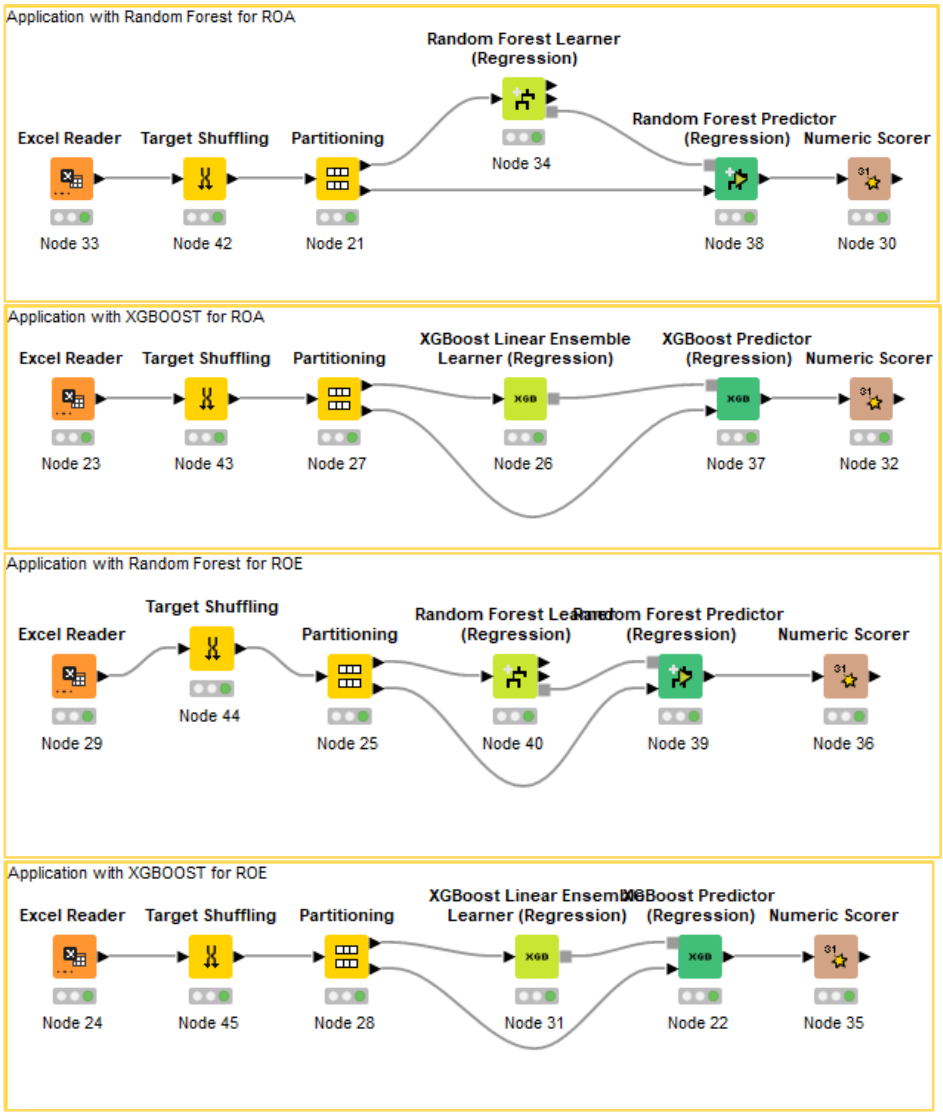


Figure 19. *The four Models after Target Shuffling.*

4.3. Findings of Multiple Linear Regression

4.3.1. ROA

A total of 248 observation is included, the results showed a very good R squared of 52.84%, which means that almost 52% of the

variation of ROA can be explained by the independent variables after taking in consideration the adjusted R squared.

Table 8. *Multiple linear regression statistics for ROA.*

Multiple R	0.73
R Square	0.53
Adjusted R Square	0.52
Observations	248

The F test showed a value of 54.22 that allows to have a significance or *P*-value of 1.26 E^{-37} which is far less than 0.05, that means that the multiple linear regression model has detected a linearity of the influence of variables, thus the regression model is valid.

Table 9. *ANOVA test results for ROA.*

	df	SS	MS	F	Significance F
Regression	5	538.94	107.79	54.22	1.26 E^{-37}
Residual	242	481.05	1.99		
Total	247	1019.99			

The linear regression model shows a P-value less than 0.05 for Capital (0.000220527), Liquidity (5.76824E^{-06}), Activity (1.01156E^{-31}), Size (0.044560132), while the Asset (0.080623908) showed a value more than 0.05, the suggested multiple linear regression for ROA:

$$\text{ROA} = 3.569 + 0.009\,(\text{Capital}) - 0.009\,(Asset) + 0.033\,(Liquidity)$$
$$-2.414\,(\text{Activity}) - 0.274\,(\text{Size})$$

Table 10. *Linear regression coefficients, t-statistic, ranking and P-value for ROA.*

	Rank	Coefficients	Standard Error	*t* Stat	*P*-value
Intercept		3.569	0.797	4.479	0.000
Capital	3	0.009	0.002	3.750	0.000
Asset	5	-0.009	0.005	-1.754	0.080
Liquidity	2	0.033	0.007	4.638	0.000
Activity	1	-2.414	0.177	-13.614	0.000
Size	4	-0.274	0.136	-2.019	0.044

This multiple regression model and according to t-statistics, suggests the ranking of the variable as Activity being the more important with t-statistic of (-13.6145), then Liquidity (4.63773), then the Capital (3.750927), then the Size (-2.01929), and finally the Asset (-1.75442) with a small probability of this influence being related to chance.

4.3.2. ROE

As in ROA, a total of 248 observation is included, the results showed a very good R squared of 79.22%, which means that almost 79% of the variation of ROE can be explained by the independent variables after taking in consideration the adjusted R squared.

Table 11. *Multiple linear regression statistics for ROE.*

Multiple R	0.89
R Square	0.79
Adjusted R Square	0.79
Standard Error	6.70
Observations	248

The F test showed a value of 184.521 that allows having a significance or P-value of $1.954E^{-80}$ which is far less than 0.05, that means that the multiple linear regression model has detected a linearity of the influence of variables, thus the regression model is valid.

Table 12. *ANOVA test results for ROE.*

	df	SS	MS	F	Significance F
Regression	5	41468.03	8293.61	184.521	$1.954\ E^{-80}$
Residual	242	10877.09	44.95		
Total	247	52345.13			

The linear regression model shows a P-value less than 0.05 for Capital (0.0009), Liquidity ($2.66E^{-10}$), Activity ($1.74E^{-78}$), Size ($8.44E^{-05}$), while the Asset (0.091105) showed a value more than 0.05 as in ROA too, the suggested multiple linear regression for ROE:

$$\text{ROE} = 8.803 - 0.039\ (\text{Capital}) + 0.040\ (Asset) + 0.227\ (Liquidity)$$
$$-23.791\ (\text{Activity}) + 2.584\ (\text{Size})$$

Table 13. *Linear regression coefficients, t-statistic, ranking and P-value for ROE.*

	Rank	Coefficients	Standard Error	t Stat	P-value
Intercept		8.803	3.789	2.323	0.021
Capital	4	-0.039	0.011	-3.362	0.001
Asset	5	0.040	0.024	1.696	0.091
Liquidity	2	0.227	0.034	6.594	0.000
Activity	1	-23.791	0.843	-28.216	0.000

Size	3	2.584	0.646	3.999	0.000

According to this regression, the tow first most important variable, and the least important variable influencing ROE are the same factors influencing ROA; the Activity with t-statistic of (-28.2159), and then Liquidity (6.594504), after those two comes the Size (3.999285), then the Capital (-3.36171), and finally the Asset (1.696352) with a small probability of this influence being related to chance.

4.4. Findings of Random Forest

The obtained metrics for RF regarding ROA and ROE are shown in table 14.

Table 14. *Random Forest's regression metrics.*

	Random Forrest	
	ROA	ROE
MSE	1.222	28.781
RMSE	1.105	5.365
R^2	0.365	0.493

The Random Forest shows an R^2 of 36.5% for ROA, it means that 36.5% of the variation in ROA, is accounted for by its regression on explanatory variables, and shows an R^2 of 49.3% for ROE, it means that 49.3% of the variation in ROE, is accounted for by its regression on explanatory variables, also Random Forest shows an RMSE of 1.105% for ROA, and 5.365% for ROE, although these value will serve to compare the model with XGBOOST model, yet this study can say

that a mean deviation of 1.105% for actual values from predicted ones of ROA generate by the model is expected, which is quite good as cited previously that a value of ROA above 5% is considered good, and above 20% is considered excellent, thus the model deviations of expected values from actual doesn't fluctuate much more than this, consequently it has very little chance to make o good ROA looks like a bad one or vis versa. But for ROE the result is different as a mean of 5.365% of deviation of expected ROE values from the actual is possible while a value of 15% and above is considered as good as cited before, thus the model gives a prediction about the value of ROE which can be used as hint, but not as good as its prediction of ROA.

Table 15. *MSE after target shuffling for each variable using Random Forest.*

Random Forrest		
	ROA	ROE
Mean squared error without shuffling (MSE)	1.222	25.965
Capital (MSEshuffling)	4.830	69.947
Capital (MSEshuffling - MSE)	3.608	41.166
Asset (MSEshuffling)	1.286	40.053
Asset (MSEshuffling - MSE)	0.064	11.272
Liquidity (MSEshuffling)	1.231	49.717
Liquidity (MSEshuffling - MSE)	0.009	20.936
Activity (MSEshuffling)	2.312	82.898
Activity (MSEshufling - MSE)	1.090	54.117
Size (MSEshufling)	3.953	45.662
Size (MSEshufling - MSE)	2.731	16.881

The RF allows to sort the variable by the highest importance for ROA, as Capital (3.608), Size (2.731), Activity (1.090), Asset (0.064), and liquidity (0.009). While for ROE, the most important variables are Activity (54.117), then comes Capital (41.166), Liquidity (20.936),

Size (16.881) and Asset (11.272). The variables sorted for ROA and ROE from the biggest MSE difference to the smallest are shown in table 16.

Table 16. *Variables sorted from biggest to smallest according to their importance as obtained by Random Forest.*

Random Forrest			
ROA		ROE	
Capital	3.608	Activity	54.117
Size	2.731	Capital	41.166
Activity	1.090	Liquidity	20.936
Asset	0.064	Size	16.881
Liquidity	0.009	Asset	11.272

4.5. Findings of XGBOOST

Obtained metrics for XGBOOST regarding ROA and ROE are shown in table 17.

Table 17. *XGBOOST regression metrics.*

XGBoost		
	ROA	ROE
MSE	1.444	49.868
RMSE	1.202	7.062
R^2	0.405	0.162

The XGBOOST shows an R^2 of 40.5% for ROA, it means that 40.5% of the variation in ROA, is accounted for by its regression on

explanatory variables, and shows an R^2 of 16.2% for ROE, it means that 16.2% of the variation in ROE, is accounted for by its regression on explanatory variables, also XGBOOST shows an RMSE of 1.202 for ROA, and 7.062 for ROE, although these value will serve to compare the model with Random Forest, yet this study shows that a mean deviation of 1.202% for actual values from predicted ones of ROA generate by the model is expected, which is quite good as cited previously that a value of ROA above 5% is considered good, and above 20% is considered excellent, thus the model deviations of expected values from actual doesn't fluctuate much more than this, consequently it has very little chance to make o good ROA looks like a bad one or vis versa. But for ROE the result is different as a mean of 7.062% of deviation of expected ROE values from the actual is possible while a value of 15% and above is considered as good as cited before, thus the model gives a prediction about the value of ROE which can be used as hint, but not as good as its prediction of ROA.

Table 18. *MSE after target shuffling for each variable using XGBOOST*

XGBOOST		
	ROA	ROE
Mean squared error without shuffling (MSE)	1.444	49.868
Capital (MSEshufling)	3.401	82.013
Capital (MSEshuffling - MSE)	1.957	32.145
Asset (MSEshuffling)	1.295	110.422
Asset (MSEshuffling - MSE)	0.149	60.554
Liquidity (MSEshuffling)	2.944	64.149
Liquidity (MSEshuffling - MSE)	1.500	14.281
Activity (MSEshuffling)	3.890	80.812
Activity (MSEshuffling - MSE)	2.446	30.944
Size (MSEshuffling)	4.663	30.921
Size (MSEshuffling - MSE)	3.219	18.947

The XGBOOST allows to sort the variable by the highest importance for ROA, as Size (3.219), Activity (2.446), Capital (1.957), liquidity (1.5), and Asset (0.149). While for ROE, the most important variables are Asset (60.554), then comes Capital (32.145), Activity (30.944), Size (18.947) and finally Liquidity (14.281). The variables sorted from the biggest MSE difference to the smallest are shown in table 19.

Table 19. *Variables sorted from biggest to smallest according to their importance as obtained by XGBOOST.*

XGBOOST			
ROA		ROE	
Size	3.219	Asset	60.554
Activity	2.446	Capital	32.145
Capital	1.957	Activity	30.944
Liquidity	1.500	Size	18.947
Asset	0.149	Liquidity	14.281

4.6. Comparison of Findings

For all 3 models, MLR, RF and XGBoost the comparison will take two aspects, first the ability of the model to detect ***correlations*** between variables, and this is possible by looking at R^2, second how good the model is in ***prediction***, and that is possible by examine MSE or RMSE.

Considering ROA, the MLR showed the highest R^2 (52.83%), then comes XGBoost with R^2 (40.5%) and lastly RF with R^2 (36.5%), which means that MLR is the better among other in detecting relationships between variables, then comes XGBoost and lastly RF. But when it comes to prediction things becomes inversed, the best prediction performance is obtained by RF with the lowest RMSE

(1.105), then comes the XGBoost with RMSE (1.202), and lastly the MLR with RMSE (1.410), just to note that the RMSE of RF and XGBoost was provided by the module regression metrics, while it was calculated for MLR from the ANOVA results table using the equation:

$$RMSE = \sqrt{\frac{SS_{\ residual}}{df_{\ residual}}}$$

In total, and for ROA, although results are somehow close, but a first conclusion could be made that the best model to detect relationship between variables is the MLR, and the best in prediction is RF.

Table 20. *Regression metrics for Random Forest and XGBoost regarding ROA.*

ROA			
	MLR	Random Forest	XGBoost
MSE	1.988	1.222	1.444
RMSE	1.410	1.105	1.202
R^2	0.528	0.365	0.405

When it comes to ROE, the MLR again exhibits the best ability to detect relationship with R^2 (79.2%), compared with RF that comes second with R^2 (49.3%) and lastly XGBoost with R^2 (16.2%). It seems that the RF succeeded to keep the best predictor place with again the lowest RMSE (5.365), compared with the MLR that performed better in predicting ROE than did with ROA with RMSE (6.704), and lastly the XGBoost with RMSE (7.062), In total, and for ROE, the best model to detect relationship is MLR, and the best in prediction is RF.

Table 21. *Regression metrics for Random Forest and XGBoost regarding ROE.*

ROE			
	MLR	Random Forest	XGBoost

MSE	44.947	28.781	49.868
RMSE	6.704	5.365	7.062
R^2	0.792	0.493	0.162

In total, RF model is more able to predict both ROA and ROE and is more able to detect correlation between ROE and the independent variables, while XGBOOST is more able to detect correlation between ROA and the independent variables.

Table 22. *The best performance models.*

	ROA	ROE
Prediction (MSE, RMSE)	1.RF 2.XGBoost 3.MLR	1.RF 2.MLR 3.XGBoost
Correlation (R^2)	1.MLR 2.XGBooST 3.RF	1.MLR 2. RF 3.XGBoost

4.7. Variables Ranking

The ranking according to MLR is based on the values of *t*-statistics exhibited in ANOVA test table, while for RF and XGBoost it is the MSE difference values (MSE after shuffling – MSE before shuffling) that indicate the importance of the variables, the tables below summarize the results.

Table 23. *MLR t-statistics for ROA and ROE.*

	ROA	ROE
Capital	3.751	-3.362
Asset	-1.754	1.697
Liquidity	4.638	6.594
Activity	-13.614	-28.216

Size	-2.019	3.999

Table 24. *MSE comparison for Random Forest and XGBOOST.*

	Random Forrest		XGBOOST	
	ROA	ROE	ROA	ROE
Mean squared error without shuffling	1.222	28.781	1.444	49.868
Capital (MSEshufling)	4.830	69.947	3.401	82.013
Capital (MSEshuffling - MSE)	3.608	41.166	1.957	32.145
Asset (MSEshufling)	1.286	40.053	1.295	110.422
Asset (MSEshuffling - MSE)	0.064	11.272	0.149	60.554
Liquidity (MSEshufling)	1.231	49.717	2.944	64.149
Liquidity (MSEshuffling - MSE)	0.009	20.936	1.500	14.281
Activity (MSEshufling)	2.312	82.898	3.89	80.812
Activity (MSEshuffling - MSE)	1.090	54.117	2.446	30.944
Size (MSEshufling)	3.953	45.662	4.663	30.921
Size (MSEshuffling - MSE)	2.731	16.881	3.219	18.947

For ROA, the MLR showed the highest *t*-test values for Activity (-13.614), then for Liquidity (4.638), after that the Capital (3.751), then the Size (-2.019) and finally the Asset (-1.754). RF detected for the Capital (3.608) compared to XGBOOST (1.957), for the Size (2.731) compared to (3.219), and for the Activity (1.09) compared to (2.446), for the Asset (0.064) compared to (0.149) and finally for Liquidity (0.009) compared to (1.500).

For ROE, the MLR showed the highest *t*-test for Activity (-28.216), then for Liquidity (6.594), after that the Size (3.999) and finally the Asset (1.697). RF detected for the Capital (41.166) compared to XGBOOST 32.145, for the Size 16.881 compared to 18.947, and for the Activity 54.117 compared to 30.944, for the Asset 11.272 compared to 60.554 and finally for Liquidity 20.936 compared to 14.281.

Table 25. *Comparison of the descending ranking obtained by MLR, RF and XGBoost for the independent variables influencing ROA and ROE.*

MLR		Random Forest		XGBoost	
ROA	ROE	ROA	ROE	ROA	ROE
Activity	Activity	Capital	Activity	Size	Size
Liquidity	Liquidity	Size	Capital	Activity	Asset
Capital	Size	Activity	Liquidity	Capital	Capital
Size	Capital	Asset	Size	Liquidity	Activity
Asset	Asset	Liquidity	Asset	Asset	Liquidity

According to MLR model, the variable that affect the most ROA and regarding t-statistics are in decreasing importance, the Activity (-13.614), Liquidity (4.638), Capital (3.751), Size (-2.019) Asset (-1.754). While the variables that affect the most ROE are in decreasing importance Activity (-28.216), Liquidity (6.594), Size (3.999), Asset (1.697).

According to RF model, the variables that affect the most ROA and regarding the difference between MSE values, are in decreasing importance, Capital (3.608), Size (2.731), Activity (1.090), Asset (0.064), Liquidity (0.009). While the variables that affect the most ROE are in decreasing importance, Activity (54.117), Capital (41.166), Liquidity (20.936), Size (16.881) and finally Asset (11.272).

According to XGBOOST model, the variables that affect the most ROA are in decreasing importance, Size (3.219), Activity (2.446), Capital (1.957), Liquidity (1.5) and Asset (0.149). While the variables that affect the most ROE are in decreasing importance, Size (18.947), Asset (60.554), Capital (32.145), Activity (30.944), and finally Liquidity (14.281).

CONCLUSION AND DISCUSSION

Bank profitability plays a crucial role in providing financial stability, high profits build up protective buffers against depressions, and involve banks in the market which restrains banks from the behavior of taking risks, more over the sources of profit matter greatly (Udaibir S. Das, 2019). The global financial crisis taught us that to understand the financial risk dynamics we need more appropriate tools than the standard approaches, these later tend to reduce data to a common data model, that doesn't take in consideration the complex, non-linear, multidimensional nature of these data, that is why the obtained results are exposed to too much distortion. Using ML algorithms allows not only to deal with the real nature of data but also reduce the processing time enormously, and thus allow to a continuous monitoring of results (Alessi, 2021).

One of the important roles of ML is classifying tasks, this has great value in terms of business applications. For example, classifying whether the bank will profit or not, study the importance of the variable's effects on bank profitability. 7 bank profitability ratios were included from 2010 to 2020 for 27 Turkish deposit banks, mainly ROA, Return ROE, Capital Adequacy, Asset Quality, Liquidity, Management Efficiency, and Bank Size.

The aim of this Study is to compare the implementation's success of MLR, Random Forest, and XGBOOST, in predicting bank profitability using bank profitability's related ratios, then using shuffling and comparing with the Mean Square Error (MSE), to judge the importance of bank-specific indicators and their effect on

profitability and rank these ratios by importance, many studies have evaluated the factors influencing bank profitability, using a method like statistical regression, and panel regression but here it is used advanced Machine Learning (ML) algorithm along with (MLR).

The up mentioned theoretical studies indicate well the importance of bank profitability, furthermore and in general they all come together when it comes to evaluate the profitability using ROA and ROE. The determinants of banks profitability are usually classified into internal and external factors. The internal determinants of banking profitability are something like the management efficacy, capital adequacy, diversification, liquidity, provisioning policy, bank size, risk of credits and overhead costs, on the opposite hand, the external determinants, can be classified as industry-related variables and macroeconomic.

These studies suggest well that there are internal and external factors that influence profitability, but to what weigh each factor exert its influence, and how much a factor is more important than other, this wasn't established well to our knowledge, the importance of this thesis is that it tries to rank many internal variables suing two state-of-the-art machine learning techniques RF, and XGBOOST.

The MLR expresses that the data has a very good linearity as *F*-test showed a significance of $1.26\ E^{-37}$ for ROA and $1.954\ E^{-80}$ for ROE, that is why the MLR performed the best in finding relationships between the variables with R^2 53% for ROA and 79% for ROE compared with RF 36.5% for ROA and 49.3% For ROE, and XGBoost with 40.5% for ROA and 16.2% for ROE, this linearity that is explained well by the MLR is the same raison that the algorithmic

regression (RF and XGBoost) came with lower results, as the flow of these algorithm is designed to deal with more non-linear distribution of variables. Although the MLR performed better in relationships, the prediction models didn't show best performance than RF for both ROA and ROE, and XGBoost for ROA, which gives more emphasis on the importance of these algorithmic regression models in prediction.

Furthermore, and relating to the negative or positive correlation of variables with profitability compared to other studies findings, considering the asset quality in this study being a risk of credit indicator, the findings in this study concerning:

- Management Efficacy results with the same findings of other study being a negatively important variable, it is negative because the less the score the better is the management efficacy as the managerial expenses are less.
- Size results coincide with one literature being negative on ROA.
- Capital Adequacy results coincide with the most literatures being positive on ROA and coincide with one literature being negative on ROE.
- Liquidity, only one literature addressed the liquidity and found it negatively related to ROA and ROE, while it is found positively related to ROA and ROE in this study.
- Risk of Credit, the majority of literatures found it positively related to ROA, some found it negative as this study found it, while for ROE, there was one literature with positive finding and one with negative while it was found positive in this study.

The table 26 summarizes the direction of correlation found in 8 literatures (L1 to L8) with the findings of this one (L9), correlation can be positive (+) or negative (-) or non-significant (NS) with L1: Athanasoglou et al. (2006), L2: Atasoy (2007). L3: Heffernan and Fu (2008), L4: Sufian and Chong (2008), L5: Flamini et al. (2009), L6: Naceur and Omran (2011), L7: Taşkın (2011), L8: Sufian (2012), L9: This Study.

Table 26. *Comparison of correlation direction findings with other literatures.*

Literature	MER		Size		CAR		Liquidity		Risk of Credit	
	ROA	ROE	ROA	ROE	ROA	ROE	ROA	ROE	ROA	ROE
L1	(-)		(+)		(+)				(-)	
L2					(+)				(+)	
L3					NS	(+)	(-)	(-)	(+)	
L4			(-)		(+)				(-)	
L5					(+)				(+)	
L6					(+)	(+)			(+)	(+)
L7	(-)	NS	NS	NS	NS	(-)			(-)	(-)
L8	(-)								(+)	
L9	(-)	(-)	(-)	(+)	(+)	(-)	(+)	(+)	(-)	(+)

As expected, the statistical metrics corresponding to ROA show more reliability compared to ROE, RMSE (1.105 Vs 5.365 for RF and 1.202 vs 7.062 for XGBOOST), this is well known in finance; the ROA reflects the profitability in more precise way than ROE as it includes the liability, furthermore, four out of five independent variables in this study (Asset, Liquidity, Capital, Bank Size) are directly related to the

Asset, which is the denominator of ROA, while no asset is included in the calculation of ROE, and as this is reflected in the metrics of RF and XGBOOST it gives a reassuring sense of the usability of these models in the prediction of financial data, this answer the first hypothesis of this study, "***The models are good in predicting profitability and correlating variables***".

Regarding the performance of RF and XGBOOST, in general the R squared are around 40% except for ROE in XGBOOST 16.2%, the RF shows for ROA and ROE 36.5% and 49.3% respectively, compared to 40.5% and 16.2% for XGBOOST which indicates that the models in general are performing with close performance, but with higher and better performance for RF, and this answer the second hypothesis of this study "**XGBoost is more efficient in predicting and correlating than RF**", to be false.

When reaching the ranking process, the difference between the highest and the lowest obtained value of MSE difference for ROA and ROE as obtained by RF (3.59 and 42.845) and as obtained by XGBOOST (3.070 and 46.273) which show a pretty large variation that allows a good ranking of these values, and this answer the third hypothesis of this study," **A ranking can be obtained**".

The table 25 summaries the ranking results for the three methods, for RF and XGBoost some of the findings are exactly the same as Size and Capital predicted for ROE, other are close like Size, Activity, Liquidity and Asset, predicted for ROA, but some are fare, like Asset, Activity and Liquidity predicted for ROE, in general the two methods show nearly similar results for ROA and less similarity for ROE.

The three model explain in average 45.75% (the total average of all R^2 obtained by the three models for ROA and ROE) of the variation of profitability being related to the five independent variables, the remaining 54.25% could be related to other macro and micro determinants not included in this study, the difference of the classification results between MLR, RF, and XGBOOST may be related to many factors:

- The size of the sample, this study ended up with 248 observations, while machine learning algorithm gives more precise results the more is fed with larger data.
- This study took only six micro determinants in consideration; five out of the six CAMELS, respectively Capital, Asset, Management, Earnings (ROA, ROE), Liquidity, and replaced the Sensitivity by the Size, while bank profitability could be affected by macro determinants too like GDP, inflation, crisis, unemployment, monetary policies in addition to other micro determinants like credit risk.
- The Psychologic influence of the disturbed economic environment on investors and depositors, and the effect of misleading news could affect the profitability.
- The period of the study has witnessed a very instable environment, with a lot of data fluctuation, which may result in different response of the algorithms.
- But more important is the fact that the internal factors can influence the bank's profitability in a bank related manner, in other words, these factors can go up the ranking or down as

stocks do, for some Banks the SIZE may play a big role while the Activity play the biggest role for other Banks, they don't affect profitability in the same order for all banks all the time, moreover, for the same bank the CAPITAL may be the most influencing factor in the two first quarters, while ASSET play this role in the two last quarters, depending on the location of branches, the social cultures, the advertising and the professionally of managers, this leads us to a more tailored ranking of internal factors that banks could monitor regularly, and report it in their internal reports.

Taking all results of the three model in consideration as votes, this study suggest that the most relevant variables affecting ROA are Activity, Capital, Size and for ROE, Activity and Size.

Regarding the impact of ML in general over macroeconomic, the simple use of ROA and ROE compared to GDP, will lead to a few lines of data, for example the sum of ROA of each bank for each year, and the same for ROE this will lead to a simple data table of 3 columns and 11 row (from 2010 to 2020), a good regression model can't be built from this data, while if more details are taken in consideration, the regression model will be more effective, only that with the traditional regression model, the more details included and the more non-linear is the relationship between variables, the harder the study will be, that is why a ML model is possibly more favourable for this kind of study, as the more data the study include, the more precise the regression model is.

At the end the main aims of this study were met, RF and

XGBOOST are good models to establish a prediction and correlation between independent variables and dependant variables in comparison with MLR, the RF algorithm showed little higher performance than XGBoost, a ranking of these independent variables was obtained and a regular changing in the ranking of these variables for banks and during the year is suggested.

For bank's managers, this Study shows well that ML algorithm could have a big help to analyse and rank the influencing factors easily and quickly, moreover, the simplicity to set a module, as it only need to be set once, and the speed of the ranking could be generated as it takes milliseconds, allows the bank to tailor their own ML algorithm to discover regularly the ranking of the influencing factors as the data being updated and decide actions upon these data and ranking.

For economist and researchers, this study suggests expanding the data for many countries and with institution of different level of financial and economic power in future studies, also including macro determinants factors like GDP, inflation, crisis, monetary policies, unemployment rates along with other CAMELS micro determinants especially sensitivity (risks), a more comprehensive model of ML could be developed in this case that offer a better predictability.

For policies makers, as ML speeds up processing data, and can spot patterns, then apply it to new data in order to predict results, ML by doing so help humans to make better decisions compared with the conventional method of making decisions. This study suggest that ML can be adopted in the monetary policy when it comes to regularize interest rates, analysing past data of the impact of rising or minimizing

interest rate on financial stability or macroeconomic, ML could predict the precise interest rate needed, or even more predict GDPs if a certain regulations are adopted, moreover policy makers should take in consideration favouring banks to use ML to monitor and submit regularly a ranking for the internal and external factors that influence the bank's profitability, these authorities can then gather all the data and detect some other possible factors that weren't known before due to the new ML capability to precisely detect any relationship from analyzing data.

REFERENCES

Alessi, L. S. (2021, January 09). Machine Learning for Financial Stability. *Data Science for Economics and Finance*, 65-87. Retrieved from https://link.springer.com/chapter/10.1007/978-3-030-66891-4_4#citeas

Al-Tamimi, H. (2010). Factors Influencing Performance of the UAE Islamic and Conventional National Banks. *Global Journal of Business Research, 4(2)*, pp. 1-9.

Ampomah, E., Qin, Z., & Nyame, G. (2020). Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement. *Information, 11*, 332.

Andriyashin, A., HHrdle, W., & Timofeev, R. (2008). Recursive portfolio selection with decision trees. *SSRN*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2894287

Atasoy, H. (2007). Expenditure-income analysis in Turkish banking sector and determinants of profitability. *Central Bank of Turkey*.

Athanasoglou, P., Brissimis, S., & Delis, M. (2008). Bank-Specific, Industry-Specific and Macroeconomic Determinants of Bank Profitability. *Journal of International Financial Markets Institutions and Money, 18 (2)*, pp. 121-136.

Athanasoglou, P., Delis, M. D., & Staikouras, C. K. (2006). Determinants of bank profitability in the South Eastern European Region. *Bank of Greece Working Paper,, 47*, pp. 1-36.

Aydın, Y. (2019, 3 25). Türk Bankacılık Sektöründe Karlılığı Etkileyen Faktörlerin Panel Veri Analizi ile İncelenmesi. *Gümüşhane Üniversitesi Sosyal Bilimler Dergisi, 10*(1), 181-189. Retrieved from https://dergipark.org.tr/en/pub/gumus/issue/44146/486202

Baele, L., De Jonghe, O., & Vander Vennet, R. (2007). Does the Stock Market Value Bank Diversification? *Journal of Banking and Finance, 31 (7)*, 1999-2023.

Beedham, M. (2019). PwC Has a New Cryptocurrency Auditing Tool… and Actually, it Sounds Alright. Retrieved from https://thenextweb.com/hardfork/2019/06/20/pwc-halotool-audit-cryptocurrency/

Berger, A. N., Klapper, L. F., & Turk-Ariss., R. (2009). Bank Competition and Financial Stability. *Journal of Financial Services Research, 35 (2)*, 99-118.

Bhargava, H. K., Sridhar, S., & Herrick, C. (1999). Beyond spreadsheets: tools for building decision support systems. *Computer, 32(3)*, 31-39. Retrieved from https://doi.org/10.1109/2.751326

Bingham, N., & Fry, J. M. (2010). *Regression Linear Models in Statistics*. London: Springer. doi:DOI 10.1007/978-1-84882-969-5

Boillet, J. (2018). How Artificial Intelligence Will Transform the Audit. *EY*. Retrieved from https://www.ey.com/en_us/assurance/how-artificial-

intelligence-will-transform-theaudit

Bossmann, J. (2016). Top 9 ethical issues in artificial intelligence. *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-inartificial-intelligence/

Breton, A. (2018). Quant Investing: The dangers of the Black Box. Retrieved from https://www.refinitiv.com/perspectives/future-of-investing-trading/quant-investingdangers-black-box/

Brignall, M. (2018). Amazon Hit with Major Data Breach Days Before Black Friday. *The Gardian*. Retrieved from https://www.theguardian.com/technology/2018/nov/21/amazon-hit-with-major-databreach-

Britannica. (n.d.). *artificial intelligence.* Retrieved from https://www.britannica.com/technology/artificial-intelligence

Budish, E., Cramton, P., & Shim, J. (2015). The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. Retrieved from https://faculty.chicagobooth.edu/eric.budish/research/HFT-FrequentBatchAuctions.pdf

Bughin, J., Seong, J., Manyika, J., Hamalainen, L., Windhagen, E., & & Hazan, E. (2019). Tackling Europe's gap in digital and AI. *McKinsey Global Institute*. Retrieved from https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-europesgap-in-digital-and-ai

Capraru, B., & Ihnatov, I. (2014). Banks' profitability in selected

central and eastern European countries,. *Procedia Economics and Finance, 16*, 587-591.

Cartea, A., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and High-Frequency trading.* Cambridge: Cambridge University Press.

CB Insights. (2019). *AI In Numbers: Global Funding, Exits, And R&D Trends In Artificial Intelligence.* Retrieved from https://www.cbinsights.com/research/report/ai-in-numbersq2-

Cecchetti, S. G., & Schoenholtz, K. L. (2015). *Money, Banking, and Financial Markets* (4 ed.). Mc Graw Hill.

Chen, J. (2019). Algorithmic Trading. *Investopedia*. Retrieved from https://www.investopedia.com/terms/a/algorithmictrading.asp

Chinner, V. (2018). Artificial Intelligence and the Future of Financial Fraud Detection. *Forbes.* Retrieved from https://www.forbes.com/sites/theyec/2018/06/04/artificialintelligence-and-the-future-of-financial-fraud-detection/#2e8ed8fb127a

Clerkie. (2019). Retrieved from https://clerkie.io

Collins. C, D. D. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda,. *International Journal of Information Management, 60*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0268401221000761#:~:text=Artificial%20Intelligence%20(A.I.)%20is%20defined,making%2C%20and%20even%20demonstrating%20creativity.

Dale, R., & & Reiter, E. (2000). *Building Natural Language Generation Systems.Series: Studies in Natural Language Processing.* Cambridge, The UK: Cambridge University Press.

Deloitte. (2018). 16 Artificial Intelligence Projects from Deloitte Practical Cases of Applied AI: Unleash the power of AI for your organization. Retrieved from https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/innovatie/deloitte-nlinnovatie-artificial-intelligence-16-practical-cases.pdf

DeYoung, R., & Torna, G. (2013). Nontraditional Banking Activities and Bank Failures During the Financial Crisis. *Journal of Financial Intermediation, 22 (3)*, 397-421.

Dietrich, D., Heller, B., & Yang, B. (2015). *Data Science & Big Data Analytics: Discovering,*. Indianapolis, The US: John Wiley &.

Dietvorst, B., Simmons, P. J., & Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them. *researchgate*. Retrieved from https://www.researchgate.net/publication/268449803_Algorithm_Aversion_People_Erroneously_Avoid_Algorithms_After_Seeing_Them_Err

Elder Reasearch. (n.d.). Target Shuffling. Retrieved from https://www.elderresearch.com/resource/innovations/target-shuffling-process/

Elena, P. (2021). Predicting the Movement Direction of OMXS30 Stock Index Using XGBoost and Sentiment Analysis. *Blekinge*

Institute of Technology.

Elsas, R., Hackethal, A., & Holzhauser, M. (2010). The Anatomy of Bank Diversification. *Journalof Banking and Finance, 34 (6)*, 1274-1287.

European Commission. (2019). Digital Single Market: Artificial Intelligence. Retrieved from https://ec.europa.eu/digital-single-market/en/artificial-intelligence

Europian Central Bank. (n.d.). Retrieved from https://www.bankingsupervision.europa.eu/banking/priorities/assetquality/html/index.en.html

Expert Systems. (2016). Natural Language Processing and Text Mining. Retrieved from https://www.expertsystem.com/natural-language-processing-and-text-mining/

EY. (2019). EY Announces the First Solution Designed to Help Gauge Impact and Trustworthiness of Artificial Intelligence Systems. Retrieved from https://www.ey.com/en_gl/news/2019/04/ey-announces-the-first-solution-designed-tohelp-gauge-impact-and-trustworthiness-of-artificial-intelligence-systems

Flamini, V. C., McDonald, & Schumacher, L. (2009). The determinants of commercial bank profitability in Sub-Saharan Africa. *IMF Working Paper*, 1-30.

Forbes. (2018). 15 Business Applications for Artificial Intelligence and Machine Learning. Retrieved from https://www.forbes.com/sites/forbestechcouncil/2018/09/27/15

-businessapplications-for-artificial-intelligence-and-machine-learning/#2190620579f2

Forum., W. E. (2015). Deep Shift Technology Tipping Points and Societal Impact. Retrieved from http://www3.weforum.org/docs/WEF_GAC15_Technological_Tipping_Points_report_2015.pdf

Ghebregiorgis. F, A. A. (2016, November 28). Measurement of bank profitability, risk and efficiency: The case of the Commercial Bank of Eritrea and Housing and Commerce Bank of Eritrea. *10(22)*, pp. 554-562.

Gul, S., Irshad, F., & Zaman, K. (2011). Factors Affecting Bank Profitability in Pakistan. *The Romanian Economic Journal, 39*, 61-87.

Gunsel, N. (2007). Financial Ratios and also the Probabilistic Prediction of failure in North Cyprus. *European Journal of research, 18(2)*, 191-200.

Hargrave, M. (2022, June 14). Return on Assets (ROA): Formula and 'Good' ROA Defined. *Investopedia*. Retrieved from https://www.investopedia.com/terms/r/returnonassets.asp#:~:text=The%20ROA%20figure%20gives%20investors,ROA%20means%20more%20asset%20efficiency.

Hayes, A. (2022, April 05). What the Capital Adequacy Ratio (CAR) Measures With Formula. *Investopedia*. Retrieved from https://www.investopedia.com/terms/c/capitaladequacyratio.asp

Heffernan, S., & Fu, M. (2008). The determinants of bank performance in China,. *Cass Business School*. Retrieved from http://ssrn.com/abstract=1247713

Hendershott, T., & & Riordan, R. (2013). Algorithmic Trading and the Market for Liquidity. *Journal of Financial and Quantitative Analysis, 48*. Retrieved from https://faculty.haas.berkeley.edu/hender/ATMonitor.pdf

Heriot-Watt University. (2017). E2E NLG Challenge. Retrieved from http://www.macs.hw.ac.uk/InteractionLab/E2E/

IBM. (n.d.). IBM Cloud Learn Hub. Retrieved from https://www.ibm.com/cloud/learn

Illowsky, B., & Dean., S. (2012). *Collaborative Statistics.* OpenStax CNX.

Javelin Strategy Report. (2015). False-Positive Card Declines Push Consumers to Abandon Issuers and Merchants. *javelinstrategy*. Retrieved from https://www.javelinstrategy.com/pressrelease/false-positive-card-declines-push-consumers-abandon-issuers-and-merchants

Joshi, N. (2019). Can AI Become Our New Cybersecurity Sheriff? *Forbes*. Retrieved from https://www.forbes.com/sites/cognitiveworld/2019/02/04/can-ai-become-our-newcybersecurity-sheriff/#11090a8836a8

Kagan, J. (2021). CAMELS Rating System: What It Is, How It Is Calculated. *Investopedia*. Retrieved from https://www.investopedia.com/terms/c/camelrating.asp#:~:text

=CAMELS%20is%20an%20international%20rating,%2C%20Liquidity%2C%20and%20Sensitivity.%22

Kaplan, A., &Haenlein, M. (2018). *Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence.* Retrieved from https://reader.elsevier.com/reader/sd/pii/S0007681318301393?token=A3915364EC5F78441839EAE336B3A051291E6195266EE7C3DFCB5672BA601F9E39126CFBC30CD1B0F24732D9321C1708

Karabulut.G. (2003, Mart 28). BANKACILIK SEKTÖRÜNDE SERMAYE KARLILIK ILIŞKISI: TÜRK BANKACILIK SISTEMI ÜZERINE BIR INCELEME. *I.Ü. Siyasal Bilgiler Fakültesi Dergisi*, 169. Retrieved from chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dergipark.org.tr/en/download/article-file/5410

Keeley, M. C. (1990). Deposit Insurance, Risk, and Market Power in Banking.". *American Economic Review, 80 (5)*, 1183-1200.

Kohler, M. (2014). Does Non-Interest Income Make Banks More Risky? Retail- Versus Investment-Oriented Banks. *Review of Financial Economics, 23 (4)*, 182-193.

Korpela, K. (2017). Big Data: A cheat sheet for the rest of us. *Medium*. Retrieved from https://medium.com/the-chic-geek/big-data-a-cheat-sheet-for-the-rest-of-us-8d64a3e5672

Krauss, C., Do, X., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on

the S&P 500. *European Journal of Operational Research, 259 (2)*, 689–702. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0377221716308657

Laeven, L. R. (2014, May). Bank Size and Systemic Risk. *IMF STAFF DISCUSSION NOTE*. Retrieved from https://www.ecb.europa.eu/events/pdf/conferences/140902/bank_size_and_systemic_risk.pdf?3a7a70d742f3995e51e4483b9fe4adbb#:~:text=Bank%20size%20is%20measured%20as,to%20total%20risk%2D%20weighted%20assets.

LeBaron, B., Brock, W., & Lakonishok J. (1992). Simple technical trading rules and the stochastic properties of stock returns. *J Finance, 47*, 1731–1764.

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support. *Decision Support Systems, 32(4)*, 361–377. Retrieved from https://doi.org/10.1016/S0167-9236(01)00121-X

Mamatzakis, E., & Remoundos, P. (2003). Determinants of Greek Commercial Banks Profitability 1989-2000. *Spoudai, 53*.

Mardanghom, R., & Sandal, H. (2019). *Artificial Intelligence in Financial Services.* Bergen: Norwegian School of Economics.

Marr, B. (2018). How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *Forbes*.

Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-createevery-day-the-mind-blowing-stats-everyone-should-read/#61c686d260ba

MasterClass. (2022). Retrieved from https://www.masterclass.com/articles/return-on-assets-guide

Mcclure, M. (2021, May 23). How ROA and ROE Give a Clear Picture of Corporate Health. *Investopedia*. Retrieved from https://www.investopedia.com/investing/roa-and-roe-give-clear-picture-corporate-health/

Medium. (2019). 7 Ways AI Can Improve Customer Experience. *ChatbotNews*. Retrieved from https://chatbotnewsdaily.com/7-ways-ai-can-improve-customer-experiencebe015f2834ba

Meiselman, B. S., Nagel, S., & Purnanandam, A. K. (2018). Judging Banks' Risk by the Profits They Report. *SSRN*. Retrieved from https://ssrn.com/abstract=3169730.

Melendez, C. (2016). Artificial Intelligence Gets into Auditing, What's Next? As AI spreads to new businesses, good software development will be crucial for achieving success. Retrieved from https://www.infoworld.com/article/3044468/artificial-intelligence-getsinto-auditing-whats-next.html

Menkveld, J. A. (2014). High-Frequency Traders and Market Structure. Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1111/fire.12038

Michal, H. (2018). Application of Machine Learning to Financial

Trading. *School of Electrical Engineering and Computer Science*.

Mitchell, C. (2019). The Two Biggest Flash Crashes of 2015. *Investopedia.* Retrieved from https://www.investopedia.com/articles/investing/011116/two-biggest-flash-crashes-2015.asp

Morde, V. (2021). "XGBoost Algorithm: Long May She Reign!". *Towards data science*. Retrieved from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Mubarak. (2019). Alarming Influence of AI and Chatbot in the Banking and Finance Industry. *mubarak*. Retrieved from https://www.mubarak.om/ai-and-chatbot-in-banking-and-finance/

Naceur, B. S., & Omran, M. (2011). The effects of bank regulations, competition, and financial reforms on banks' performance. *Emerging Markets Review, 12*, 1-20.

Naceur, S. (2003). The Determinants of the Tunisian industry Profitability: Panel Evidence. *University Libre de Tunis*.

Natalya, M., Ratnovski, L., & Vlahu, R. (2015). Bank Profitability and Risk-Taking. *IMF Working Paper*, 15/249.

Nisbet. R, G. D. (2017). *Handbook of Statistical Analysis and Data Mining Applications.*

Öz, B., Ayriçay, Y., & Kalkan.G. (2011). FİNANSAL ORANLARLA HİSSE SENEDİ GETİRİLERİNİN TAHMİNİ: İMKB 30 ENDEKSİ HİSSE SENETLERİ ÜZERİNE DİSKRİMİNANT ANALİZİ İLE BİR UYGULAMA. *ANADOLU ÜNİVERSİTESİ SOSYAL BİLİMLER DERGİSİ, 11*(3), 51-64. Retrieved from https://earsiv.anadolu.edu.tr/xmlui/handle/11421/205

P.Sarkar. (2022, November 28). What is Regression Analysis? Types, Techniques, Examples. *knowledge hut*. Retrieved from https://www.knowledgehut.com/blog/data-science/regression-analysis-and-its-techniques-in-data-science

Panetta, K. (2017). Neural Networks and Modern BI Platforms Will Evolve Data and. *Gartner*. Retrieved from https://www.gartner.com/smarterwithgartner/nueralnetworks-and-modern-bi-platforms-will-evolve-data-and-analytics/

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 (1),* 259-268.

Petria, N., Capraru, B., & Ihnatov, I. (2015). Determinants of banks' profitability: evidence from EU 27 banking systems. *Procedia Economics and Finance, 20,* 518-524.

Prep Waal Street. (n.d.). Retrieved from https://www.wallstreetprep.com/knowledge/return-on-equity-roe/

R. G. Miller. (1997). *Beyond ANOVA: Basics of Applied Statistics.* Chapman & Hall.

Rani, D. M., & Zergaw, L. N. (2017). Bank specific, industry-specific and macroeconomic determinants of bank profitability in Ethiopia. *International Journal of Advanced Research in Management and Social Sciences, 6(3),* 74-96.

Rasiah, D. (2010). Theoretical Framework of Profitability as Applied to Commercial Banks in Malaysia. *European Journal of Economics, 19,* pp. 74-97.

Ray, S. (2018). Four Types of AI. *Medium.* Retrieved from https://codeburst.io/four-types-of-ai-6aab2ce57c19

Reis, Ş. G. (2016). Banka Karlılığını Etkileyen Faktörler: Türkiye Örneği. *Muhasebe ve Finansman Dergisi, 72,* 21-36. Retrieved from https://dergipark.org.tr/en/pub/mufad/issue/35669/396715

Rençber, Ö. F., & Yücekaya, P. (2021). GİRİŞİMCİLERİN YABANCI ÜLKELERE YATIRIM KARARLARI AÇISINDAN İŞ YAPMA KOLAYLIĞINI ETKİLEYEN FAKTÖRLER ÜZERİNE BİR ARAŞTIRMA. *EUropean Journal of Managerial Research*, 206-220. Retrieved from https://dergipark.org.tr/en/pub/eujmr/issue/59611/880668

Reynoso, R. (2019). 4 Main Types of Artificial Intelligence. *Learning Hub.* Retrieved from https://learn.g2.com/types-of-artificial-intelligence

Rohner, P., & & Uhl, W. M. (2017). Robo-Advisors versus Traditional

Investment Advisors: An Unequal Game. *The Journal of Wealth Management*. Retrieved from https://doi.org/10.3905/jwm.2018.21.1.044

Roman, A. (2013, December). Analysing the Financial Soundness of the Commercial Banks in Romania: An Approach Based on the Camels Framework. *Procedia Economics and Finance, 6*, 703-712. Retrieved from https://www.researchgate.net/publication/258507030_Analysing_the_Financial_Soundness_of_the_Commercial_Banks_in_Romania_An_Approach_Based_on_the_Camels_Framework

Rongyuan, Q. (2022). The Construction of Corporate Financial Management Risk Model Based on XGBoost Algorithm. *Hindawi Journal of Mathematics*.

Rouse, M. (2010). Electronic discovery (e-discovery or ediscovery). Retrieved from https://searchfinancialsecurity.techtarget.com/definition/electronic-discovery

Rouse, M., & Stedman, C. (2018). Text Mining (Text Analytics). Retrieved from https://searchbusinessanalytics.techtarget.com/definition/text-mining

San, O. T., & Heng, T. B. (2013, February 28). Factors affecting the profitability of Malaysian commercial banks. *African Journal of Business Management*, 649-660. Retrieved from http://www.academicjournals.org/AJBM

Schmidhuber, J. (2014). Deep Learning in Neural Networks An Overview. *sciencedirect*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0893608014002135?via%3Dihub

Schmidhuber, J. (2015). Deep Learning. *Scholarpedia*, 32832.

Schmidhuber, J. (2015). Deep Learning. *Scholarpedia.* Retrieved from http://www.scholarpedia.org/article/Deep_Learning

Senator, E. T., Goldberg, G. H., Wooton, J., Cottini, A. M., Khan, U., Klinger, D. C., . . . Wong, R. (1995). The FinCEN Artificial Intelligence System: Identifying Potential Money from Reports of Large Cash Transactions. Retrieved from https://www.aaai.org/Papers/IAAI/1995/IAAI95-015.pdf

Seth, S. (2019). The World of High-Frequency Algorithmic Trading. *Investopedia.* . Retrieved from Seth, S. (2019). The World of High-Frequency Algorithmic Trading. Investopedia. Available at: https://www.investopedia.com/articles/investing/091615/world-high- frequencyalgorithmic-trading.asp

Shaher, T., Kasawneh, O., & Salem, R. (2011). The Major Factors that Affect Banks' Performance in Middle Eastern Countries. *Journal of Money, Investment and Banking, 20*, pp. 01-109.

Silipo, R. (2019). From a Single Decision Tree to a Random Forest. *Towards Data Science*. Retrieved from https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

Sobowale, J. (2016). Beyond Imagination: How artificial intelligence is transforming the legal profession. *ABA Journal, 4*. Retrieved from http://www.abajournal.com/magazine/article/how_artificial_intelligence_is_transforming_the_legal

Son, H. (2017). JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours. *Bloomberg.* Retrieved from https://www.bloomberg.com/news/articles/2017-02-28/jpmorgan-marshals-an-army-of-developers-to-automate-high-finance

Sorensen, E., Miller, K., & Ooi, C. (2000). The Decision Tree Approach to Stock Selection. *The Journal of Portfolio Management, 27 (1),* 42–52. Retrieved from https://jpm.pm-research.com/content/27/1/42

Spacey, J. (2018, March 11). 6 Examples of Management Efficiency. Retrieved from https://simplicable.com/new/management-efficiency

Sufian, F. (2012). Determinants of bank profitability in developing economies: empirical evidence from the South Asian banking sectors. *Contemporary South Asia, 20 (3)*, 375-399.

Sufian, F., & Chong, R. R. (2008). Determinants of bank profitability in a developing economy: empirical evidence from Philippines. *Asian Academy of Management Journal of Accounting and Finance, 4(2)*, 91-112.

Taşkın, F. D. (2011). (2011), The factors affecting the performance of

the Turkish commercial banks,. *Ege Akademik Review,, 11 (2)*, 289-298.

The Economic Times. (n.d.). Retrieved from https://economictimes.indiatimes.com/definition/return-on-equity

Thompson, C. (2014). *Cybercrime Costs Global Economy $400 Billion.* CNBC. Retrieved from https://www.cnbc.com/2014/06/09/cybercrime-costs-global-economy-

Topak, M. S., & Talu, N. H. (2016). Internal Determinants Of Bank Profitability: Evidience From Turkish Banking Sector. *International Journal of Economic Perspectives*, 37-49. Retrieved from http://www.econ-society.org

Tuovila, A. .. (2021). Overall Liquidity Ratio. *Investopedia*. Retrieved from https://www.investopedia.com/terms/o/overall-liquidity-ratio.asp

Turing, M. A. (1950). Computing Machinery and Intelligence. *Mind*, pp. 433-460. Retrieved from https://www.csee.umbc.edu/courses/471/papers/turing.pdf

Udaibir S. Das, K. H. (2019, April 30). Bank Profitability: Consider the Source. *IMF BLOG*. Retrieved from https://www.imf.org/en/Blogs/Articles/2019/04/30/blog-bank-profitability-consider-the-source

Van Liebergen, B. (2017). Machine Learning: A revolution in Risk management and Compliance? Retrieved from

https://ideas.repec.org/a/ris/jofitr/1592.html

Weisbach, S. M., Tan, C., Stern, H. L., & Erel, I. (2019). Selecting Directors Using Machine Learning. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3144080

Xu, Y., Li, Z., & Luo, L. (2013). A Study on Feature Selection for the Trend Prediction of Stock Trading Price. *International Conference on Computational and Information Sciences*, 579-582. Retrieved from https://ieeexplore.ieee.org/document/6643074

Yaninen, D. (2017). Artificial Intelligence and the Accounting Profession in 2030. Retrieved from https://cpapng.org.pg/data/documents/CPA-Presentation-Artificial-Intelligenceand-the-Accounting-Profession-in-2030_1.pdf

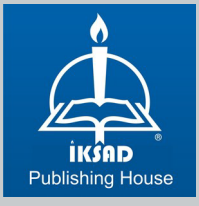
Yetgin, R., & Ekşi, İ. H. (2017). KOBİ'lere Kredi Verme Tutumu: Türk Bankacılık Sektöründe Bir Uygulama. *Business and Economics Research Journal*, *8*(3), 487. Retrieved from https://www.proquest.com/docview/1933856199?pq-origsite=gscholar&fromopenview=true

Zheng, T., Ziqin, Y., & Guangwei, Z. (2019). Stock selection with random forest: An exploitation of excess return in the Chinese stock market. *Heliyon*, *5*(8). Retrieved from https://doi.org/10.1016/j.heliyon.2019.e02310.

Zhu, M., Philpotts, D., & Stevenson, M. (2012). The benefits of tree-based models for stock selection. *Journal of Asset*

Management, 13 (6), 437-448. Retrieved from http://link.springer.com/10.1057/jam.2012.17

Zou, Z. B., Peng, H., & Luo, L. K. (2015). The Application of Random Forest in Finance. *Applied Mechanics and Materials, 740*, 947–951. Retrieved from https://doi.org/10.4028/www.scientific.net/amm.740.947



ISBN: 978-625-367-157-0