Faculty of Economics and Business Administration

Machine Learning for the Explanation and Prediction of M&A in the Energy and Utilities Sector

Author: Lucía Ramos Fitera
Director: María Coronado Vaca

MADRID | June 2023

Table of Contents

ABSTRACT

The research paper presented uses five different machine learning techniques to explain and predict M&A targets in the Energy and Utilities Sector across the seven continents (in 85 countries) for a period of time of ten years, between 2013 and 2022. To do so, the study uses a dataset of 1471 public companies, of which 1239 are non-targets and 232 have been targets previously.

First, it aims to provide investors with the necessary information to determine which and how financial variables affect target selection of companies operating in the selected industry by using a Logit explanatory model. Second, it analyses which machine learning model out of five possible options (Logistic Regression, KNN, Random Forest, Ensemble with Logit as the stacking model, and Ensemble with Decision tree as the stacking model) is the best performer in predicting M&A targets in the mentioned sector, to help investors make profitable investment decisions.

These objectives answer an identified gap in the existing literature, and therefore have significant importance. Previous research has been done to find out which financial variables determine M&A, or to determine the optimal predictive model for M&A targets. We have observed that the literature focuses on using Machine Learning models like Logit, Random Forest or Ensembles in order to predict targets, but without being focused on a particular sector, focusing on an alternative sector as ours, or focusing on a particular country. Furthermore, all of the studies have used a lower number of models in comparison to our analysis. Because of this, none of these have been focused solely on the Energy and Utilities sector, highly relevant in today's economic context. Moreover, no study has compared such a high number of models as ours, making our comparative unique.

The study concludes that the most important financial factor that determines whether a company in the Energy and Utilities sector will become target is profitability, with acquirers looking for companies with high EBITDAs but lower EBITs, thus targets being characterised by having high depreciations. Furthermore, the best performing model identified above all was the Logit stacking model, followed by Random Forest as the best performer in terms of AUC.

KEY-WORDS

Energy, Utilities, M&A, Target, Acquirer, Machine Learning, Prediction, Explanation, Financial Variable.

RESUMEN

El presente trabajo utiliza cinco técnicas de aprendizaje automático para explicar y predecir compañías objetivo de fusiones y adquisiciones en el sector de la Energía y Utilidades dentro de los siete continentes (en 85 países) para un periodo de tiempo de 10 años, entre 2013 y 2022. Para ello, se utiliza una base de datos con 1471 compañías públicas, de las cuales 1239 no han sido objetivo y 232 sí lo han sido en el pasado.

En primer lugar, el estudio tiene como objetivo aportar a los inversores la información necesaria para determinar qué y cómo las variables financieras afectan a la selección de compañías objetivo que operan en el mencionado sector utilizando un modelo *Logit* explicativo. En segundo lugar, analiza qué técnica de aprendizaje automático entre cinco posibles opciones (Regresión Logística, *KNN, Random Forest, Logit Ensemble stacking model,* y *Decision tree Ensemble stacking model)* tiene el mejor rendimiento en predecir compañías objetivo en el sector seleccionado, con el fin de ayudar a los inversores a tomar decisiones de inversión rentables.

Estos objetivos responden a una brecha en la literatura existente, y por lo tanto tiene gran importancia. Otros estudios se han realizado para determinar qué variables financieras afectan a las fusiones y adquisiciones, o para determinar cuál es el modelo óptimo para predecir compañías objetivo. Hemos visto que se han enfocado en utilizar modelos de *Machine Learning*, como *Logit*, *Random forest*, o *Ensembles* para predecir compañías objetivo, pero sin centrarse en un sector concreto, centrándose en un sector alternativo al nuestro o centrándose en un país concreto, y además siempre utilizando un número inferior de modelos que nuestro análisis. Es por ello, que ninguna de estas obras analizadas se ha enfocado específicamente en el sector de la Energía y las Utilidades, una industria muy relevante en el contexto económico actual. Además, ninguno ha comparado un número tan elevado de modelos como nosotros, siendo nuestra comparativa única.

El estudio concluye que el factor financiero más importante que determina si una compañía en el sector de la Energía y las Utilidades será objetivo es la rentabilidad. Los adquirientes buscan compañías con *EBITDAs* altos peros *EBITs* bajos, y por ello las compañías objetivo se caracterizan por tener una depreciación alta. Además, el modelo con mejor rendimiento identificado ha sido el *Logit stacking model,* seguido por *Random Forest* como el mejor en términos de *AUC.*

PALABRAS CLAVE

Energía, Utilidades, Fusiones y Adquisiciones, Compañía objetivo, Compañía adquiriente, Aprendizaje Automático, Predicción, Explicativo, Variable Financiera

FIGURES INDEX

TABLES INDEX

1. INTRODUCTION

1.1 OBJECTIVES

The following study has been developed around two main objectives. First, the study wants to analyze which financial variables have the most influence to determine whether a company is classified as a target or not in M&A in the Energy and Utilities sector. Second, it has the objective to find out which Machine Learning model is the most precise when determining whether a company in the Energy and Utilities sector will be an M&A target or not.

1.2 MOTIVATION FOR THE STUDY

M&A have become essential in today's economy, because of its input provided to returns for both investors and company owners. They are a key part in creating shareholder value, helping companies grow in a fast way, and making sure worse performing companies are rescued and combined with more successful ones (Tamosiuniene & Duksaite, 2009).

M&A success leads to stock markets performing better, and therefore returns for investors increasing. We saw this clearly on the recent M&A boom experienced in 2021. KPMG (2021) explains that "Global mergers and acquisition activity in 2021 easily surpassed the pre-pandemic level and nearly matched the peaks of 2015 and 2007". This significant raise in M&A activity in 2021 was led mainly because of several macroeconomic factors, which include low interest rates set by central banks and the economic recovery after Covid. M&A transactions in 2021 amounted to more than 5 trillion dollars (KPMG, 2021). This significant raise in M&A activity influenced the stock market, with it achieving historical maximums and making stock increase its value.

We therefore consider the study to be of singular importance to investors who want to obtain returns in the investment markets. For this reason, the research has been developed, in order to offer investors the necessary information to know in which metrics they should focus on when choosing a company in which to invest. If they choose to invest in a company that will be acquired in the future it is highly likely that they will be able to make a profit from the investment. Kolostyak (2021) from *Morningstar* explains that when an M&A transaction is about to take place the acquirer's share price usually decreases, while the target's share price tends to go up.

The fact that we are currently in a year characterized by turmoil, with high volatility and uncertainty, due to several recent events such as the Covid Crisis, the Ukraine war, high inflation, and the collapse of Silicon Valley Bank, reinforces the importance for investors to be able to know which investments will be profitable, as they want to protect themselves from loses in crisis periods.

Furthermore, the research paper is focused solely on the energy and utilities sector as it is a sector with special importance. Not only energy-related corporates but also private equity funds currently search for investments in energy businesses that are willing to contribute to the energy transition. Furthermore, given the recent energy crisis, the current macroeconomic environment presents an opportunity for M&A in the sector given the unbalanced energy supply and demand.

By carrying out the study and through our literature reviews, we have observed that there are many papers that focus on M&A predictions and its determinants. However, there is a lack of studies which are centred around this field specifically in the Energy sector. Given the importance of the sector explained above, this study will therefore contribute additional value to the existing literature.

Finally, we have seen that there are many studies that focus on M&A prediction using logit analysis. Although we initially believe this model can be effective in making predictions, we also want to challenge the idea. Because of this, a comparison of models is carried out in our research, in order to determine which one is the most precise when predicting targets involved in M&A deals in the Energy sector.

1.3 **METHODOLOGY**

The methodology used to answer our two objectives is divided in two parts. First, an explanatory model using logit regression has been developed in order to answer our first objective of which financial variables determine M&A targets in the Energy and Utilities Sector. Second, five machine learning models (logit regression, KNN, random forest, and two ensembles) have been used in order to determine which of the five is the most accurate in making predictions of M&A targets in the Energy and Utilities sector. In order to select the best performing model, several metrics have been used, including accuracy, AUC, classification error, type I error, sensibility, and specificity.

Both parts have been developed using R as the programming language in R studio. The dataset used was obtained from FactSet and consists of 1471 observations, of which 1239 are non-targets and 232 have been targets in the past. Companies were public, headquartered around the 7 continents (in 85 countries) and operated in one of the 5 subsectors chosen (Electric Utilities', 'Oil & Gas Production', 'Integrated Oil', 'Gas Distributors', 'Coal', 'Alternative Power Generation', 'Oil Refining/Marketing', and 'Water Utilities'). The part of the dataset of companies that had been target in the past was for a time frame of 10 years (2013 to 2022).

8 financial variables were used as explanatory or predictive variables, and this include 'Enterprise Value', 'Revenue', 'EBIT', 'EBITDA', 'Total Assets', 'Long Term Debt', Cash and Short-Term Investments', and 'Price to Earnings'. The foundation behind using these variables and not others

lies on the fact that we have carried out a literature review of financial determinants in order to find out which variables to include in our study. Furthermore, since our machine learning problem was a classification problem, our dependent variable could take values of either 'target' or 'non-target'.

Finally, apart from the source of data used from FactSet to carry out our analysis, we have also based our study on different sources to gather as many insights as possible on the M&A Energy Sector and the determination of targets. In order to do so, we have made an extensive review of literature of different academic papers, high quality reports from well-known consultancy businesses, reliable website articles from professionals, recent newspaper articles and educational websites.

1.4 STRUCTURE

The paper has been structured in the following way. First, a theoretical framework is given so that the reader is able to obtain knowledge on the theory behind the study. This part contains information on what we understand from M&A, what is the energy and utilities sector, trends of M&A in the energy and utilities sector, what financial variables are considered important determinants of M&A, and finally a literature review on the different studies other researchers have carried out in the past that give insights on which Machine Learning models are the most precise in determining M&A targets.

In section 3, a description of the empirical study used to carry out our analysis is presented. In this section, a description of the dataset and of the methodology used are included, where the reader can understand why the data was chosen, how it was processed, and the different models, steps and performance measures used to gather the results presented in the paper.

The study continues with section 4, where we include the different findings obtained from our analysis and how we answer our investigation questions. This part includes the results from the explanatory model to determine the most relevant financial variables to determine M&A in the Energy and Utilities sector, as well as a comparison between the performance measures of different machine learning models.

Finally, a section on conclusions, a bibliography, and an annex of the code used in R are presented in section 5, section 6, and section 7 respectively.

2. THEORETICAL FRAMEWORK

2.1 INTRODUCTION TO M&A

Since our research paper is focused on transactions known as "Mergers and Acquisitions", we would first like to start our theoretical framework on what we understand from these concepts, and why they are important, not only in the financial industry, but for the whole economy.

When we talk about a merger or an acquisition, we are talking about the creation of a new entity from the combination of two companies (Roberts et al., 2003). The CFA Institute (n.d) defines an acquisition as the process by which a company purchases either a majority or minority stake of another company. On the other hand, a merger involves one company being absorbed by another, dissolving either only the absorbed company, or both, to create a new entity in what is known as consolidation. Roberts et al. (2003) explain that the difference between the two concepts arises from how the combination takes place. While a merger often occurs because of negotiations between the two parties involved, an acquisition is normally when a company buys another, either in a friendly or a hostile way.

Malik et al. (2014) in their research paper *Mergers and acquisitions: A conceptual review* give a brief introduction on why companies engage in M&A. The authors explain that all organizations have a common goal, which is maximizing returns in order to be able to give incentives to shareholders and maximize their wealth. There are several ways of increasing profits for a company, and one of them is engaging in Mergers and Acquisitions. Mergers and Acquisitions allow to increase the size of a company, develop new products, increase customer base, or operate in other geographical areas in a fast and efficient way. They are an alternative to organic growth, which requires more time, and a way to allow small firms which lack of financial resources to keep operating in the market by merging or being acquired by the larger players. Any firm that decides to participate in M&A has the objective to operate more efficiently with others than on an alone basis.

If we go further into the motives on why companies do mergers and acquisitions by reading what Leepsa & Mishra (2016) present in their research paper *Theory and practice of mergers and acquisitions: Empirical evidence from Indian cases*, we see that motives for participation in these types of deals are unique and vary from one transaction to another. Basing the paper on the different theories of M&A and carrying out a literature review, the paper identifies the following reasons for engaging in mergers and acquisitions: in order to increase efficiency, in order to achieve synergies, for diversification purposes, for strategic realignment, to benefit from market inefficiencies where targets are undervalued, for information purposes, due to agency problems, to acquire free cash flow, to increase market power, for tax considerations and finally to redistribute wealth. Surprisingly we can see there are infinite reasons on why a company will be

seeking to carry out M&A. The study also suggests that M&A can result in creating, destroying, or maintaining value of companies.

Synergies are the reason behind the highest number of M&A transactions (Leepsa & Mirsha, 2016). We can define a synergy as an increase in value, meaning that the sum of the individual values of the acquirer and the target participating in an M&A transaction is less than the value of both entities operating together: Value [Single entity formed by Acquirer and Target] > Value [Acquirer] + Value [Target] (Feldman & Hernandez, 2022). In the article *A McKinsey perspective on value creation and synergies* by McKinsey (2010), they identify three types of synergies that can be achieved. These are cost, capital, and revenue synergies. Cost synergies involve reducing costs by becoming more efficient and less redundant. Capital synergies consist of managing the balance sheet better, by improving metrics such as working capital, assets, or debt. Finally, revenue synergies arise from growth due to cross-selling, increasing customer base or being able to offer new business lines. Whilst cost synergies are easier to achieve, revenue synergies tend to be more unrealistic and optimistic. Furthermore, DePamphilis (2003) in its book *Acquisitions and Other Restructuring Activities: An Integrated Approach to Process, Tools, Cases, and Solutions* divides the types of synergies between operational and financial synergies. The author says operating synergies come from the need for having a better operating efficiency by incorporating economies of scale or scope, or through the acquisition of complementary assets or skills. In contrast, financial services are those that come from the possibility of the acquirer reducing its cost of capital.

From this section we learn that there are many reasons on why companies take part in M&A and see how important they are for companies in order to grow and generate returns. We want also to point out that from now onwards, in our research paper we will not make a differentiation between the terms merger and acquisition. We thought it was important to make a distinction in this section for clarification purposes, but we will now use both terms interchangeably, as this is a common practice in financial studies to achieve simplification.

2.2 INTRODUCTION TO THE ENERGY AND UTILITIES SECTOR

Now that we know what M&A transactions are, we want to clarify what we mean by the second specification of our field of study: The energy and utilities sector. Whilst many claim they know what types of companies this sector compromises, we want to ensure our readers comprehend in which businesses our study will focus on.

The MSCI (2023), in its definitions of Global Industry Classification Sectors (GICS), define the energy sector as the sector which “comprises companies engaged in exploration & production,

refining & marketing, and storage & transportation of oil & gas and coal & consumable fuels. It also includes companies that offer oil & gas equipment and services." Also, they define the utilities sector as the one that "comprises utility companies such as electric, gas and water utilities. It also includes independent power producers & energy traders and companies that engage in generation and distribution of electricity using renewable sources". We understand from these definitions that both of these sectors constitute companies that are involved in the generation, transportation and storage of different kinds of energy, including electricity, therefore from now onwards we will use the term "energy" to describe both the energy and utilities sector.

The MET Group (2022), a Switzerland based European integrated company of energy, explains in its article *Energy Sector Definition: How does the energy industry work?* the different parts that constitute the energy sector and the functioning of it. The article suggests the industry can be divided between fossil fuels and renewable energy sources. The first involve traditional energy sources, such as oil, natural gas, and coal. These sources, although being efficient in generating energy, are considered very harmful for the environment and it is predicted that we will run out of them in the medium term. In contrast, renewable energy sources include more environmentally friendly generation processes but dependant on weather conditions. These include different types of power: hydro, solar, wind, geothermal and biomass.

From the explanations from MET Group (2022) it is easy to misunderstand that companies in the "Energy sector" as defined by the MSCI tend to be fossil fuel related, whilst the ones in the "Utilities sector" as per the MSCI are more related to renewable energies. However, this is not true, since now, more than then, we are seeing how traditional oil and gas companies are moving towards more sustainable investments and using capital to become involved in the energy transition.

Companies in the energy and utilities sector are more and more pushing towards being part of the energy transition. S&P Global (2020) in its article of *What is Energy Transition?* clarifies that energy transition is the process of shifting from fossil fuel sources to renewable sources. The research also mentions that the increase in the transition is being led by a growing interest of investors in ESG factors and climate risks. The energy transition is not only affecting the generation of energy, but also transmission and storage infrastructures.

2.3 M&A IN THE ENERGY AND UTILITIES SECTOR

After defining what M&A and the energy and utilities sector are, we are now going to look at some of the trends currently occurring around this field across the globe.

In Bain & Company's (2023) *Global M&A Report 2023,* we get an overview on what has happened recently and what is expected from M&A involving energy companies and the energy transition. Almost a third of every deal occurring in the energy sector is related to energy transition, and in the first three quarters of 2022, spin offs in the energy sector amounted to $250 billion. Companies in the mentioned industry are more and more increasing their focus on removing brown assets from their portfolios and replacing them with greener assets.

M&A deals in the sector are being influenced by several factors. The high amount of cash in energy companies' balance sheets and the increasing need for more renewable energy is driving M&A transactions. In the first 9 months of 2022, 27% of total deals in the energy sector were related to the energy transition, and growing (Bain & Company, 2023). Even though traditional oil and gas companies usually use their free cash flows to pay dividends to shareholders, there is now an outlook for them to engage in deals led by different drivers. These include not only energy transition purposes, but also as a way to secure licenses, increase productivity, improve cost efficiency and build new skills (Deloitte, 2023).

In the research paper *Developments and Trends of Mergers and Acquisitions in the Energy Industry* by Andriuškevičius & Štreimikienė (2021) the authors carry out a PESTEL (political, economic, social, technological, environmental, and legal) analysis to determine which factors drive the M&A industry in energy companies. The authors conclude that both volume and deal value, as well as deciding to engage in M&A and if the deal will be successful are all affected by PESTEL. In particular, M&A in the energy sector is being influenced by prices of commodities, oil supply and demand, renewable energy growth trends, and regulation. The M&A market we are focused on is cyclical and has tendencies that move like waves depending on the macroeconomic factors that surround them.

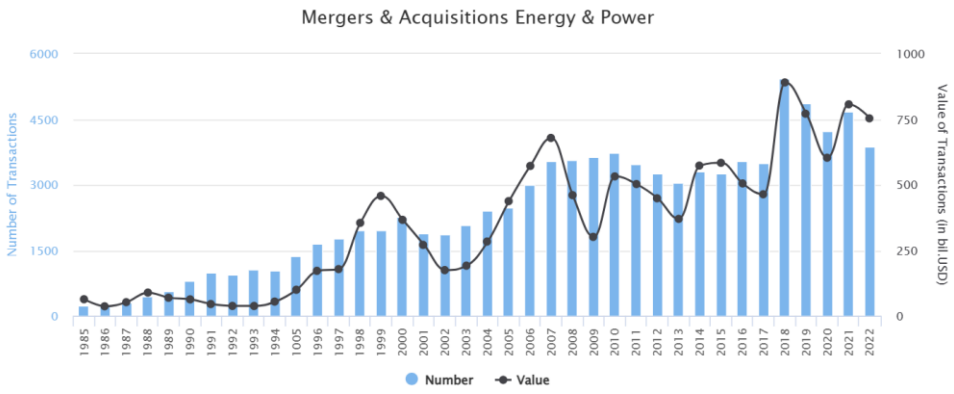
In the newspaper article by Wilson (2022) from the *Financial Times* titled *Oil and gas majors: time for a transformative clean energy deal?* we can read several examples of big oil companies that are starting to move into the renewables sector. The article emphasises again how the big oil ang gas companies have large amounts of cash due to the soaring of prices of oil in 2022, and that this could be an opportunity for them to start acquiring renewable companies. However, the problem currently is that big oil companies are engaging in smaller deals, and not transformational ones, mainly due to valuations of companies that would make the deal dilutive. The challenge for oil and gas companies targeting renewable companies is to generate returns from the deals performed and to be able to run the acquired businesses more efficiently than the targets would have done on a stand-alone basis. Because of this, according to Wilson (2022), big oil companies will keep using their excess cash for stock buybacks and paying dividends, and perform deals at a smaller scale, in the $1bn-$5bn range.

Focusing on the renewables sector, we read what Mckinsey (2022) writes about M&A for renewable developers. The article mentions that there is a strong appetite for M&A related to renewable energy as shown by the high trading multiples of this type of companies. Deal value was of $425 million in the first two quarters of 2022, a significant increase from the average of $150 million in 2018, as McKinsey (2022) explains. Because of this increased interest, competition is high, achieving successful M&A is becoming challenging and acquirers need to follow a disciplined approach on order to generate value from the deals. Not only independent renewable companies are becoming engaged in these deals, but also oil and gas majors and financial institutions are getting involved in order to enhance new skills and keep up with trends. This vast amount of competition is creating difficulties in being cost effective when acquiring a company.

McKinsey (2022) also talks about which geographical areas are standing out in renewables M&A. While Europe has been the leader in deal activity for the last four years, North America is starting to catch up with substantial growth. The Pacific, Asia, and the rest of the continents are slightly behind, but growing too (McKinsey, 2022).

Finally, we have observed what has been the historical evolution in number and deal value for the Energy & Power sectors from the data provided by the IMAA Institute. As we can observe from **Figure 1,** from 1985 to 2022 there has been an upward trend in terms of both the number of transactions and value of transactions. In 2018, numbers reached historical maximums with values of 5,439 deals and $890 billion deal value. From that point onwards, there was a decrease due to the Covid 19 in 2020, but this trend rapidly scaled up to reach high values in 2021, when M&A was experiencing one of its best years. We can also see from the trend in the graph that mergers and acquisitions in the sector are cyclical, with decreases in number and value during turmoil periods, but with an overall rising trend in the markets.

Figure 1. M&A number of deals and deal value from 1985 to 2022 in the Energy and Power Sector



Source: IMAA Analysis. IMAA Institute

From this section we can conclude an outlook for mergers and acquisitions in the energy and utilities sectors. They are likely going to increase in the coming years, considering that trends will be affected by macroeconomic factors. Both big oil companies and pure renewable businesses are highly interested in engaging in deals in order to grow and benefit from synergies. However, this high demand creates competition in the market, making it difficult to perform deals that are accretive in value.

2.4 **FINANCIAL DETERMINANTS OF M&A**

Even though in the previous section we have mentioned some of the specific drivers of M&A activity in the energy and utilities sector, in this part of our paper we are going to focus solely on the financial drivers that tend to increase the likelihood of an M&A transaction. Since our methodology involves analysing how financial metrics determine if a company will be considered a target or not, we wanted to comprehend what other literature says about these metrics.

First, the M&A Research Centre at Cass Business School et al. (2016) present in their paper a study on what the financial characteristics of acquisition targets are. The research paper identifies six metrics that can be used to predict if a company will become a target or not in M&A activity. Furthermore, the article divides their findings between public and private companies. Since our study is focused on public companies, we thought it was relevant to collect the results from these. The authors say that "Public companies are more likely to become acquisition targets if they are small, fast-growing, with low profitability, low leverage, low liquidity and low valuations". Moreover, the most important factors that determine whether a public company will become an acquisition target are size and profitability. A smaller and less profitable public company will likely be a target. The rationale behind this is that public companies tend to have big sizes, therefore, to fit the acquirer they should be smaller than average. For profitability, the reasoning behind looking for low profits are that this implies lower valuations and therefore lower prices. They are a good deal as long as they have the ability to grow.

KPMG (2007) along with Professor Steven Kaplan from the University of Chicago Graduate School of Business in their work *The Determinants of M&A Success* also give insight on some of the characteristics that make a good target. Along with other variables, the paper analyses what the P/E ratio of the target should be. Findings show that returns for acquirers are higher if the P/E ratios of targets are lower than average. The authors explain that this is because lower P/E ratios mean prices of targets are fairer and imply more realistic cash flows.

Kapil and Dhingra (2021) carry out a literature review from different authors of some of the different determinants involved in Mergers and Acquisitions. The study divides the factors involved in several sections, but we have focused our attention on the factors that are firm specific.

One of the factors that influence M&A is the net debt capacity, in other words, the amount of excess cash of a company or how healthy a company is based on its debt (Bruner, 1998, Hunter et al., 2000, as cited in Kapil & Dhingra, 2021).The market value at which a target company is trading and overall financial performance also influence M&A activity (Laamanen, 2007, as cited in Kapil & Dhingra, 2021). Other authors state that smaller targets that are not financially profitable are also considered targets in deals by financially capable acquirers. (Danzon et al., 2007, as cited in Kapil & Dhingra, 2021).

Moreover, in the article from Osbornet et al. (2012) titled *The preferences of private equity investors in selecting target acquisitions*, we can learn about some of the factors taken into account by private equity firms in order to select targets. We are aware our study is focused on strategic acquirers buying other firms instead of financial acquirers like private equity houses, but we found it interesting to also include information on what this type of acquirers search, to provide the reader with further knowledge on the topic. Osborne et al. (2012) conclude that private equity investors tend to focus more on firm specific factors rather than macroeconomic factors. The firm specific factors that indicate in which company a financial buyer will bid on include it being less volatile in terms of trading price on the stock market, having the capacity to grow and being of a reasonable size.

Finally, when we analyse the paper *Mergers and Acquisitions: expected vs actual performance. A set of case study assessments* by Sebastiano (2021), we learn about the due diligence process an acquirer follows when selecting a target and in which factors they focus on. Sebastiano (2021) indicates that in order to take a decision, first the possibility to generate synergies is considered. This can be observed through the business plans of the target. Second, several financial metrics of the potential target are analysed. This include if the business is healthy in terms of capital structure, potential for growth and its debt. Share price should also be considered, by determining whether the firm is trading at higher or lower values than it should be. Targets should be a good fit if operations are good and can generate synergies, but also if debt ratio levels are appropriate and if the target is going to be able to refinance any debt involved in the deal. Because of this, looking at cash flow levels is also important. Sebastiano (2021) also indicates in his paper that P/E and Earnings Per Share ratio are also a determinant. His study indicates that acquirers feel the creation of value will be easier after the deal if the targets P/E is lower and EPS is higher. This way, the combined companies will be able to trade at higher levels after the deal and thus create value. Furthermore, the target having high R&D expenses could mean that it has high growth prospects and therefore become an interesting target.

This section concludes that the number of factors that can determine M&A is infinite. Therefore, studies in the field are necessary in order to add to the existent literature more prove on how M&A can be predicted. For this reason, our study could have a significant impact.

2.5 LITERATURE REVIEW OF PAST STUDIES

In this final section, we have gathered together the results of different academic papers that focus on the prediction of M&A using several machine learning techniques. The objective of this section is to collect opinions given by authors so that the reader is able to see how past studies have carried out the research and what results have been obtained, regarding which models work best in making predictions.

First, Furenmo (2020) in his thesis *Predicting Corporate Takeover Outcomes Using Machine Learning*, compares what machine learning technique is the best approach in order to predict M&A deal outcomes. The author compares the performance of two models: Logit and Random Forest, by using a training set of 5922 deals and a test set of 1481 observations obtained from Bloomberg, both for transactions taking place from 2000 to 2018. The variables used for the model include how much time the deal took to complete, whether there were rivalry bids involved, what the total deal value was, how was the payment form (i.e. stock or cash), whether the deal for friendly or hostile, how much bid premium was paid, how much leverage was used, what was the percentage ownership targeted, and finally whether the buyer was strategic or financial. The paper concludes that the random forest model is a better predictor than the logit model in terms of prediction accuracy, specificity, and sensitivity, reinforcing the idea that other machine learning techniques should be sought in order to analyse which are better predictors. Furthermore, the author explains that target size and whether the attitude of management was positive or negative influenced the outcome of the deal.

Kim & Arbel (1998) also carried out a logit model used to identify which variables seemed significant in order to predict merger targets in the hospitality sector in the period of 1980 to 1992. The study presents several hypotheses on 9 variables in order to identify which if this are significant by using a dataset of 69 companies that were merger targets and 192 companies that were not merger targets in the sectors of restaurants, hotels without gaming and hotels with gaming facilities. The results show that four variables seem to be relevant in predicting targets. A company which will likely be an M&A target is characterised by being substantially large, experiencing a difference between its liquid resources and its growth prospects, having a high capex to total assets ratio and finally a low price to book ratio meaning the firm is undervalued (Kim & Arbel, 1998). The study also concluded that the logit model was an efficient tool in order

to identify which companies could become M&A targets. However, the authors indicate that the tool should be used as support, but that other techniques should be considered in order to make a clearer decision.

In the paper *Predicting Australian Takeover Targets: a Logit Analysis* by Peat & Stevenson (2008), the authors use financial statement variables in order to predict M&A targets using several different regression models and comparing the results with economic criteria to analyse chance, with the objective to be able to generate returns for investors if models are better at classifying targets than chance. Data is obtained for the period between 1995 and 2006 from the financial statements of Australian public companies. Results from the paper show that the model that has better predictive accuracy is a combined adjusted model, based on industry adjusted financial ratios, and that using this model to make predictions and therefore identifying in which companies to invest instead of being led by chance, can generate higher returns for investors. Finally, the paper also concludes that although the logit model worked well, it is advisable to combine models in order to have better forecasts.

The Nicholas Center for Corporate Finance & Investment Banking (2019) carry out a novel study in predicting M&A targets using high quality machine learning models with predictive power. These include random forest, neural network and ensemble models, being the ensemble model the most powerful one in predicting targets across US public companies. The ensemble model was composed by a combination of the individual neural network and random forest and identified the 10 companies that were mostly likely to be acquired in the following 12 months. Again, the authors indicate that although predicting power being successful, the model cannot be used alone and should be combined with other types of analysis. This other analysis includes fundamental analysis, meaning taking into account who owns the company, information on management, business lines of the company and metrics that are specific for the industry (Nicholas Center for Corporate & Investment Banking, 2019). The paper also reinforces the idea that there is a data limitation on the available information to be used in the models and that M&A in listed companies is rare, therefore limiting the study and allow space for further research.

Finally, Aramyan (2021) carries out a study presented in the academic article *Predicting M&A Targets Using Machine Learning Techniques* in which the author builds a predictive model to identify targets involved in M&A and then analyse whether these predictions could produce returns for investors. The author uses Refinitiv Data to extract 656 target companies in the US in the period of April 2010 to June 2021, and then collects information of peers of these targets in order to construct the non-target group. The methodology involves two logistic regression models, one with intermediary clustering in terms of liquidity and leverage ratios, and one without clustering to identify whether this affects results. Regarding the variables that affect results in the

model without clustering, the study identifies that companies that are most likely to be acquired have inefficient management, low prices to sales ratios, higher debt to EV, and generate abnormal positive returns. For the clustering model with companies with higher liquidity and lower leverage, the results are similar, except that companies with lower debt to EV ratios are the ones that are likely to be acquired. The study finally concludes that a logistic model that involves clustering is able to make more accurate predictions, and that a portfolio compromised of these predictions can generate abnormal returns for investors.

We can conclude from our literature review that the logit model is an accurate model in order to predict M&A targets and therefore, its efficiency should be tested. However, according to previous studies, other models have better prediction capabilities, such as combining logit with clustering, or using other machine learning techniques like random forest, neural networks, or ensembled models. What authors also agree on is that the models cannot be used on a standalone basis and that they should be combined with other techniques in order to achieve better results, and thus be able to generate returns for investors. This proves the necessity of our research. We want to compare which models can carry out better predictions of M&A targets in the Energy Sector, where there is a lack of research as we have identified through our literature review.

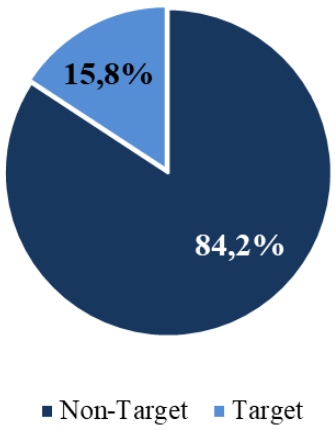
3. EMPIRICAL STUDY

The empirical study we have carried out is focused on the two objectives of our paper. First, we carry out an explanatory model using logistic regression in order to analyse which variables are the most significant in the prediction of M&A Targets in the Energy & Utilities Sector. Second, we compare five different models to determine which one is the most effective in making predictions between logit, KNN, random forest, and two ensembles of logit, decision trees and KNN.

3.1 DATASET

In this section we are going to explain the dataset we have used to carry out our study. The data has been downloaded from FactSet and consists of 1471 observations, of which 1239 are non-targets and 232 have been targets in the past. In order to filter the data, we took public companies headquartered around the 7 continents, concretely in 85 countries. The principal sectors of the companies were the 'Energy minerals' and 'Utilities' sectors, and this included the subsectors 'Electric Utilities', 'Oil & Gas Production', 'Integrated Oil', 'Gas Distributors', 'Coal', 'Alternative Power Generation', 'Oil Refining/Marketing', and 'Water Utilities'. For the subsection of target companies, this were companies that had been targets in completed M&A processes in the past 10 years, from 2013 to 2022.

Figure 2. Pie Chart of Target and Non-Target distribution



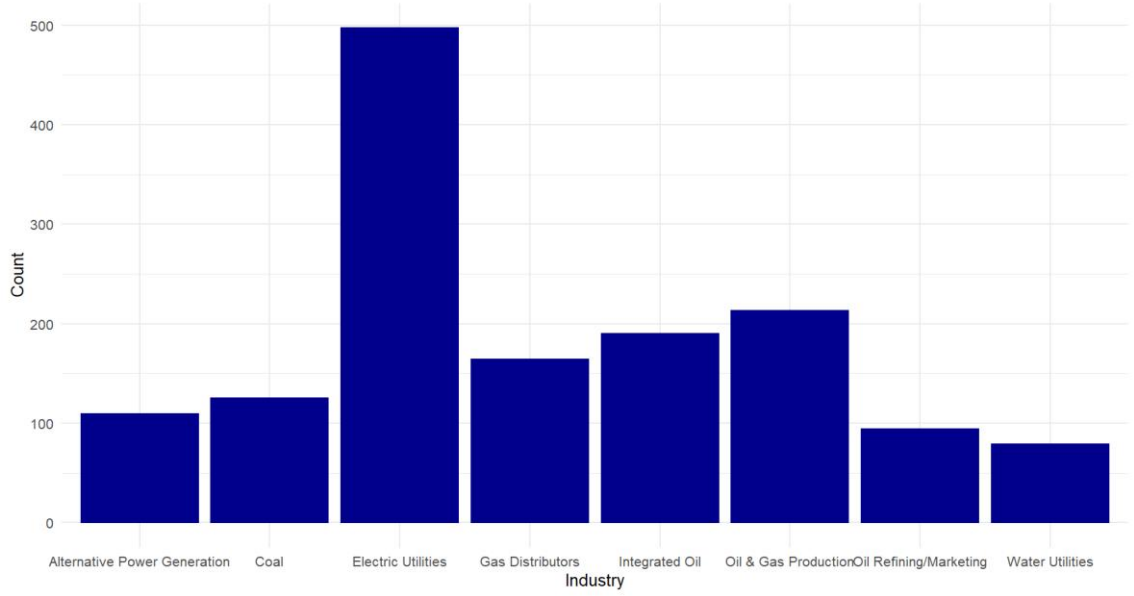
Source: Own elaboration from FactSet data

The reason why we chose only public companies was due to the availability of the data. Private companies don't tend to disclose information publicly; therefore, we thought that in order to have available data points we needed to focus only on public companies. This was a limitation of our study and an improvement that could be considered in the future.

Also, the reason why we chose a time frame of 10 years for the companies that were targets was that we wanted to consider the most recent deals possible. As it is true that the time frame could have been shorter in order to reduce the effect of business cycles, narrowing it to 10 years was acceptable given again the lack of information of deals available.

As we can see in **Figure 3** and **Table 1,** the majority of companies included in our dataset were from the 'Electric Utilities' subsector, followed by 'Oil and Gas Production' and 'Integrated Oil', representing c. 60% of our dataset.

Figure 3. Number of companies included per sector in the dataset



Source: Own elaboration using FactSet data

Table 1. Frequency of companies included per sector in our dataset

Industry	Frequency	%
Electric Utilities	497	33,8%
Oil & Gas Production	213	14,5%
Integrated Oil	190	12,9%
Gas Distributors	164	11,1%
Coal	125	8,5%
Alternative Power Generation	109	7,4%
Oil Refining/Marketing	94	6,4%
Water Utilities	79	5,4%

Source: Own Elaboration from FactSet Data

In terms of countries, out of the 85 countries included in the dataset, nearly 20% of companies were headquartered in the US, with the next two most frequent countries being China and Canada, as observed in **Table 2** where we can see the number of companies for the top 10 countries.

Table 2. Frequency of companies included per country in our dataset for the top 10 countries

Country	Frequency	%
United States	267	18,2%
China	156	10,6%
Canada	128	8,7%
Russian Federation	62	4,2%
India	56	3,8%
United Kingdom	54	3,7%
Australia	48	3,3%
Brazil	48	3,3%
Hong Kong	41	2,8%
Thailand	40	2,7%

Source: Own Elaboration from FactSet Data

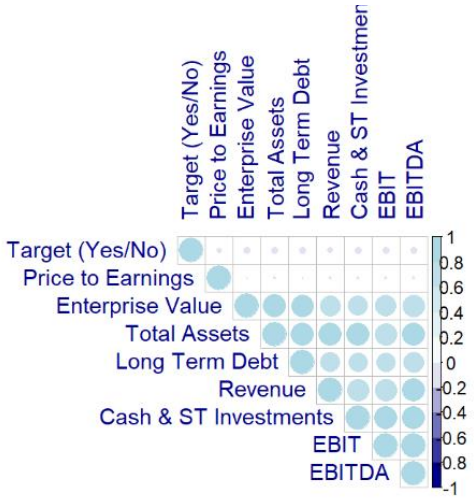
The theoretical framework presented in the correspondent section presents a basis in order to select the financial variables we have used. This were all denominated in Euros despite their reporting being in a different currency, in order to ensure comparison between observations. Furthermore, they represent either the value at the moment the company was acquired or merged in case of target companies, or the most recent financial reporting available for the non-target companies. The 8 financial variables chosen for the study were the following:

1. Enterprise Value: This variable shows the fair value of the company in the market. We thought this was important to be included as it shows the size of the company. We want to study if having a larger size affects positively or negatively if the company will become a target or not.
2. Revenue: shows the quantity of sales of the company. We want to see whether having higher sales affects being a target or not.
3. EBIT: Shows the operating profit of the company. This value is important, especially for capex intensive companies like the ones we are including in the study, because it doesn't include depreciation. Depreciation can be used as a proxy for capex; therefore, EBIT considers capex as an expense.
4. EBITDA: shows the operating profit of the company, adding depreciation, to remove how capex affects whether a company will become a target or not. It is used commonly as a proxy in finance for free cash flow.
5. Total Assets: this variable is used to analyse whether an asset-heavy company is more or less probable of becoming a target.
6. Long Term Debt: this variable is used to analyse if high leverage can affect the likeliness of being target, as having higher debt implies higher risk.

7. Cash and Short-Term Investments: the liquidity of a company is also considered, to find out whether acquirers prefer highly liquid or less liquid companies as targets.
8. Price to Earnings: this ratio can be used to determine whether a company is undervalued or overvalued in the market. We want to find out if acquirers tend to target companies that are overpriced or under-priced.

We have also decided to do a numerical analysis of the 8 financial variables explained above, in order to find out how these are correlated with each other and with the target variable. To do so, we have created a correlation matrix, as shown in **Figure 4.** This figure shows the relationship in terms of correlation between all numeric variables in our dataset and our target variable. We can observe that our binary dependent variable is not correlated with any of the 8 independent variables, supposing no problem for our study. If there were any variable correlated with our target variable, then we would need to eliminate them to improve precision. However, it is true that we can see that, apart from P/E ratio, the rest of financial variables for our observations are correlated with each other, since a stronger colour means stronger correlation. Because of the nature of our study and the way financial reporting works, it is no surprise that these variables depend on one another, as they are part of, or derived from, the financial statements of a company, which are all linked in one way on another. Because of this reason, we decided to maintain the 8 variables in our study.

Figure 4. Correlations between Target variable and financial variables of our model

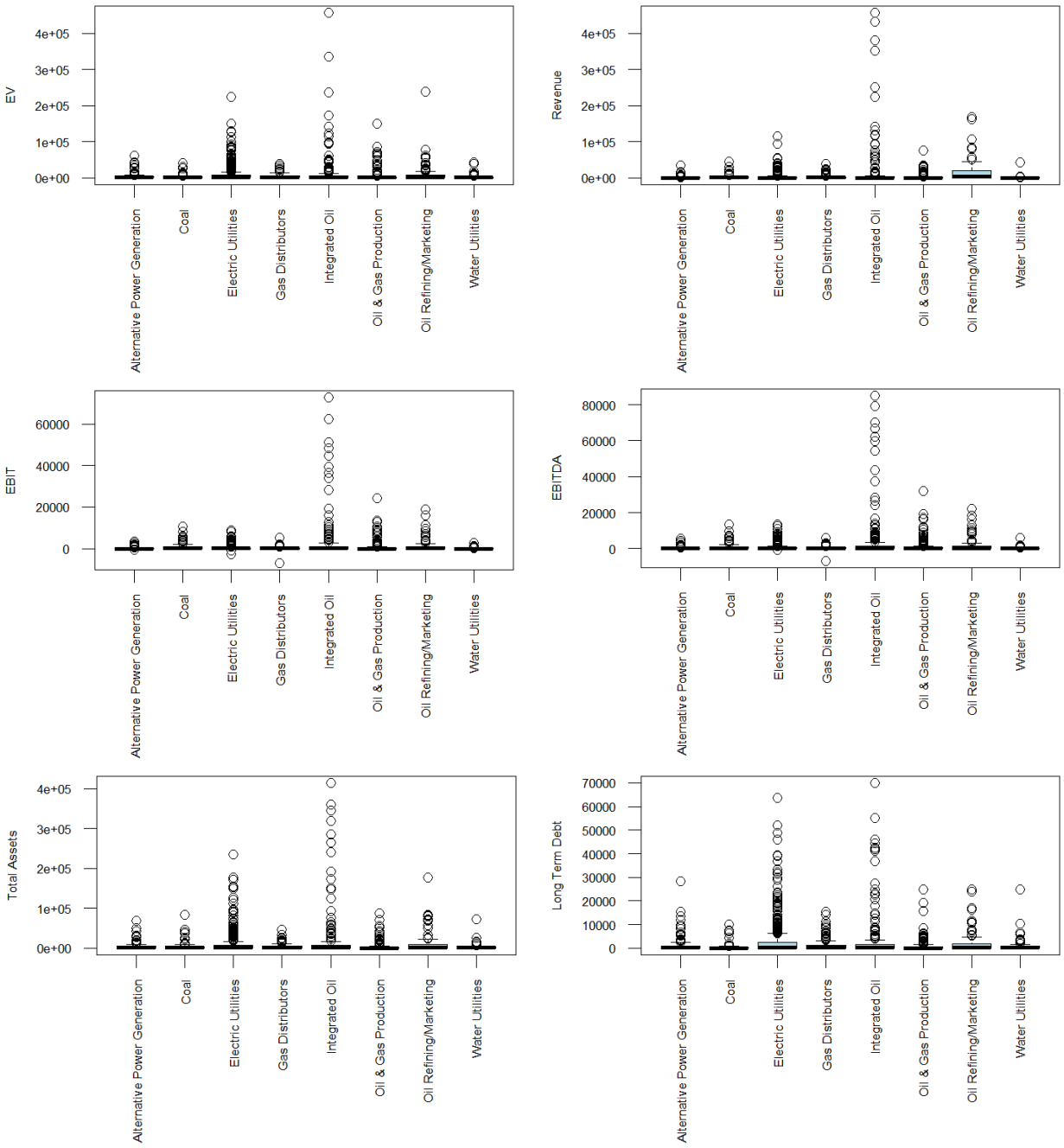


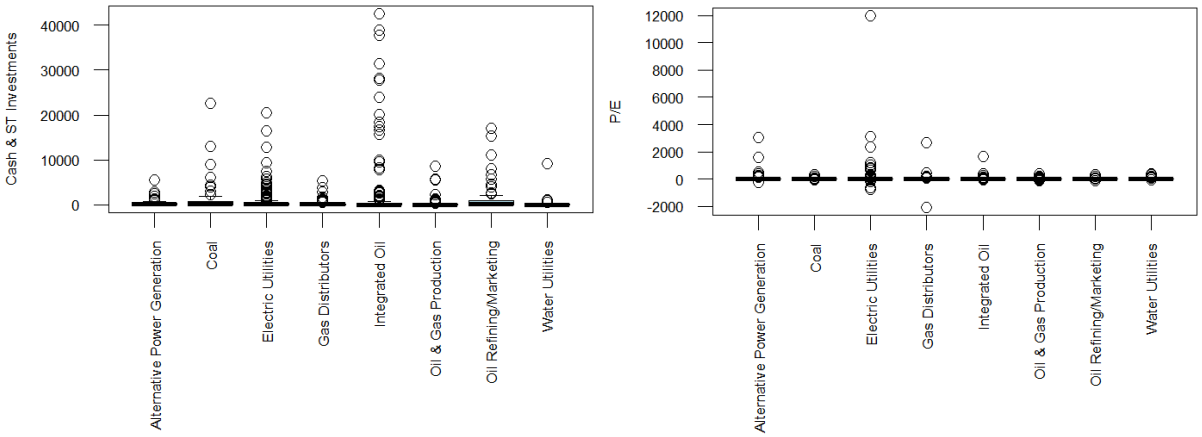
Source: Own elaboration from FactSet Data

Finally, we have created a boxplot for each of the financial variables, divided by subsector, in order to see both the magnitude and the dispersion for each of the independent variables. This is show in **Figure 5.** We can conclude first that the 'Integrated oil subsector' is the one that presents

most dispersion in the data for all variables, followed by 'Electric Utilities' for EV, Total Assets, and P/E. Second, that the financial variable that represents most dispersion across all subsectors is Long Term Debt, meaning that companies in the dataset have very varied values of Long-Term Debt compared to one another, implying very different risk levels. Finally, we can see that each of the financial variables has a different scale. This is not surprising as the magnitude for P/E must be much smaller than for EV as an example. Because of this, data should be standardised in order to achieve accurate results. This is further explained in our methodology section.

Figure 5. Boxplot per financial variable across Energy and Utilities Sectors





Source: Own elaboration using FactSet data

3.2 METHODOLOGY AND MODEL CREATION

In this section, we are going to go in detail through the approach that has been followed in order to do our study, explaining which models have been used and how they have been developed. The programming language used for our study was R using R studio. The objective of this section is to provide the reader sufficient information to understand where results come from in our next section.

First, it is important to know that all the 8 financial variables explained in the dataset section will be used in our model as independent variables. The dependent variable used in our study will be "Target (Yes/No)", a binary variable which will take values between 0 and 1, depending on whether the company is non-target or target respectively.

Before going through the models that have been chosen, it is worth knowing how the data was processed in order to prepare it for use. Once it was downloaded from FactSet using the filters mentioned in our Dataset section, we decided to go through a few steps to fine tune them. First, the target variable was converted into a factor variable since this way it was considered as a binary variable that could be classified. Second, we analysed whether there were any missing values in our data set, but the result obtained was that the dataset was complete.

We thought that removing outliers was not necessary in our dataset. Removing outliers depends massively on the objective of the study and nature of data. Normally we would delete outliers if the data was incorrect and was not a true representation of the population in order to make the study more precise. However, in our case, since the dataset is from an established provider (FactSet) we assumed data points must be correct. Also, although we saw there were some extreme values, since we are working with companies of different sizes, the fact that differences between maximum and minimum values was large for each of the financial variables was normal. We did not want to remove outliers as this could affect our results and reduce the effectiveness of our analysis. Finally, since our dataset does not have a substantially large number of observations

due to the lack of public financial information, eliminating extreme values would reduce the dataset to a very small size and worsen our results.

Instead of removing outliers, we took the approach of standardizing our financial variables, to make them of a similar size. Standardizing variables is a process in which variables are transformed to have zero mean and standard deviation of 1. This is a key step when using logistic models to explain and predict, because of the way the logistic function works in relating the dependent variable with the explicative or predictive variables. If we omitted this step, results would not have been valid, therefore the standardization of variables is a key process before continuing our analysis.

Once we had our dataset fully processed and prepared, we divided our analysis in two parts to answer our two objectives:

1. Explanatory model: In this part, we use a logistic regression model using the 8 financial variables to identify the significance of each variable in explaining whether the company is a target or a non-target. We also analyze the marginal effect each of the variables has on the target variable, to see the effect of a one unit increase in each variable has on the dependent variable: Target (Yes/No).
2. Predictive model: This part involves the comparison of predictive results of different machine learning models to answer the question of which model is the best that can be used to predict whether a company will become target or not. We have chosen here to compare logistic regression, KNN, random forest and two ensembles of logit, decision trees and KNN. For this section, we decided to partition our dataset randomly between test and training subsets, using 70% of our data as the training set, and 30% as our test set, as this are the standards normally used in machine learning studies. The training set will be used to train our models, and the test set will be used in order to test the predictive capabilities of our models.

3.2.1 EXPLANATORY MODEL

In order to explain the effect each of the financial variables we have chosen to analyze whether a company is classified as a target or non-target have, we decided using logit regression was the best option. The reason for choosing this model, and not another one, such as a simple linear regression, lies on the mathematical calculation. Since our target variable is binary, we want to force it to take values of 0 or 1. Also, we want to be able to interpret what is the probability of a company being classified as a target depending on what the value of the explanatory variables are, and the logit model provides the possibility to do so through marginal effects. Finally, we

want to be able to see what the importance of each of the variable is, or in other words, the significance, in determining whether our dependent variable will take a value of 0 or 1.

The probability of a dependent variable belonging to one category, or another, given different independent variables can be modelled using logistic regression (Furenmo, 2020). Stock & Watson (2015) in its book *Introduction to Econometrics*, give an overview on what a logit regression model consists of. They provide the following formula for the logit regression model:

$$PR(Y=1 \mid X_1, X_2, \ldots, X_K) = \frac{1}{1+e^{-(\beta 0+ \beta 1x1+\beta 2x2+\cdots+\beta kxk)}} \quad \textbf{(1)}$$

In this model, the cumulative distribution function to calculate probabilities on whether the dependent variable takes a value of 1 is the cumulative standard logistic distribution function (Stock & Watson, 2015).

The above formula applied to our study is the following:

$$PR(Y=\text{Target} \mid X_1, X_2, \ldots, X_8) = \frac{1}{1+e^{-(\beta 0+ \beta 1x1+\beta 2x2+\cdots+\beta kx8)}} \quad \textbf{(2)}$$

Where Y is our dependent variable (Target (Yes/No)) taking values in the range [0;1] and X_1, X_2, …, X_8 are our financial variables Enterprise Value, Revenue, EBIT, EBITDA, Total Assets, Long Term Debt, Cash and Short Term Equivalents, and Price to Earnings respectively.

One thing to note when interpreting results from a logistic regression as an explanatory model is that the *betas* don't have a direct interpretation. In other words, their values don't show the effect each financial variable will have on our target variable due to the nature of the formula. For this reason, average marginal effects need to be calculated, as individually they depend on the value the explanatory variable takes. This is because of the "S" shape of the logit regression line due to its exponential formula, where the value of Y = Target depends not only on the beta, but also on the value of the explanatory value X.

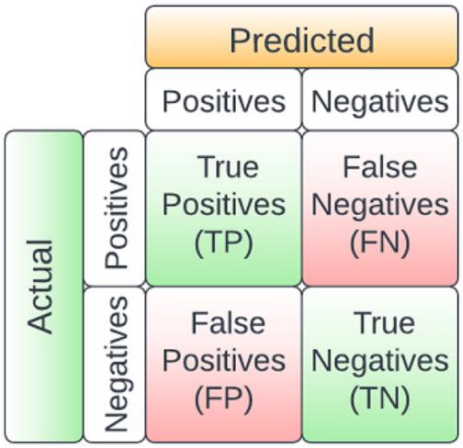
3.2.2 MEASURES TO CHOOSE BETWEEN PREDICTIVE MODELS

Also, we are going to give some theoretical background on what metrics we chose to decide between the different predictive models. We thought that the most relevant measures that were going to be used to identify the best model were a confusion matrix showing accuracy, specificity and sensibility, and the AUC metric along with the ROC curve.

First, we obtained the confusion matrix for each of our tested models. As Kundu (2022a) in his article explains, the confusion matrix is commonly used to test how good a binary classification model is in predicting outcomes. The matrix groups the classification results observed with the model in four different classes: true positive, true negative, false positive, and false negative.

While a true positive and a true negative is an observation that has been classified correctly as positive or negative, a false positive refers to an observation that is negative but has been classified as positive. Vice versa occurs with a false negative (Kundu, 2022a). An overview of how a confusion matrix looks like can be seen in **Figure 6.**

Figure 6. Structure of a confusion matrix



Source: Skevofylakas, M. (2022, July 12). *Modelling and Evaluation – Artificial Intelligence regression and classification performance metrics.* Refinitiv.

The matrix calculates the sum of observations for each of the different four terms and allows other metrics to be calculated that we include below:

- Accuracy: measures how precise in identifying cases correctly the model is.

$$\frac{Sum\ of\ true\ positives\ and\ true\ negatives}{Sum\ of\ all\ cases} \quad \textbf{(3)}$$

- Sensitivity: measures the capacity to identify positive cases correctly.

$$\frac{Sum\ of\ true\ positives}{Sum\ of\ true\ positives\ and\ false\ negatives} \quad \textbf{(4)}$$

- Specificity: measures the capacity to identify negative cases correctly.

$$\frac{Sum\ of\ true\ negatives}{Sum\ of\ true\ negatives\ and\ false\ positives} \quad \textbf{(5)}$$

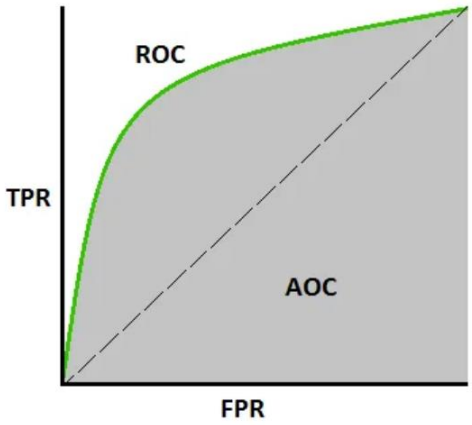
In our case, positive means being classified as a target ("1"), while negative means being classified as a non-target ("0"). Since the reasoning behind our study focuses on helping investors put their money in companies that will become targets, the worst error is to have a false positive, where we give the prediction of a company being classified as target but in reality, it is not. This is because we assume it is better for investors not to make a gain than to risk having a loss. In

other words, it is better for an investor to lose an investment opportunity (false negative) rather than doing the wrong investment (false positive). Because of this, we prefer a model that has a lower error in terms of false positives. We therefore need to focus on developing a model with higher specificity, as we want to avoid having false positives.

The second measurement used to see visually which model is the most precise in making predictions is the AUC and the ROC curve. The Area Under the Curve (AUC) and Receiver Operating Characteristics (ROC) Curve measure the performance of classification models with different decision thresholds to obtain the true positive and the false positive rates (Narkhede, 2018).

Narkhede (2018) explains that the ROC is a graphical depiction of a curve that plots the true positive rate (or sensitivity) against the false positive rate (or 1-specificity), as shown in **Figure 7.** The curve is drawn by changing the decision threshold (or cut off) and plotting the different sensibility and specificity values for each cut off. The best curve is the one that is closer to the left top of the graph, while the worst is the one that is closer to the diagonal line. This is because the further away the line of the graph, the higher the proportion of true positives.

Figure 7. Depiction of the ROC Curve



Source: Narkhede, S. (2018, June 26). *Understanding AUC – ROC Curve.* Towards Data Science.

The AUC is also a metric used to measure the prediction performance. It "represents the degree or measure of separability" and "tells how much the model is capable of distinguishing between classes" (Narkhede, 2018). Because of this, a higher AUC means a better predictive model, and can take values between 0,5 and 1, being 1 a perfect predictor.

3.2.3 PREDICTIVE MODELS

To finalise our methodology and model creation section, we are going to give a brief theoretical overview of the different models used in the predictive part of in our study in order to make the

understanding of the next section of results easier. Also, we are going to explain why we chose these models.

1) Logistic regression

The first model we chose to include to make our predictions was our logit model used as explanatory including the 8 financial variables with no interactions. We thought it was interesting to use this model because it has been widely used in past studies to predict M&A targets and therefore believe it is important to use it as a benchmark with more complex models. The theory behind it has already been explained in the explanatory section. Since this type of model is used to model the probability of our public Energy and Utilities companies belonging to the "target" or "non-target" categories, we need to set a cut-off or threshold to classify our observations as "positive" or "negative" ("target" or "non-target" in our case respectively). It is common to use 0,5 as a cut-off in machine learning studies, however, we want to let the reader know that the cut off chosen to make predictions using logit was 0,7. This means that out of 10 companies, 3 will be classified as "target". Every company that surpasses this cut off will be classified as "target". We chose a slightly higher cut-off because we wanted to prioritise having less false positives, as we thought this was our worst error, as explained above.

2) KNN

KNN is the second supervised learning technique we are going to use in order to predict our classification of companies. The method classifies observations based on their similarity with its "K nearest neighbours". Because of this, our model uses cross validation in order to obtain the optimum number of "K" and make classification predictions based on this parameter, by classifying each observation as "target" or "non-target" based on the most frequent class across the nearest observations. We have chosen to use this model because it is easy to understand and does not require a lot of computational time given the size of our dataset. Furthermore, it is well-known for having good performance and is commonly used across machine learning projects. As with Logit, we have chosen to use all 8 financial variables in our KNN model.

Raschka (2018) indicates that this model stores training observations with a class label, and then this is processed during the prediction phase. Because of this, the algorithm is classified as "lazy". Furthermore, the model is "instance-based" because it is based on comparing observations in the test set with observations in the training set, instead of with a global dataset (Raschka, 2018).

3) Random Forest

Before explaining what our third model chosen, random forest, is, it is important to know first what type of model a decision tree is. Liberman (2017) explains in its article that a decision tree is a model that consists of a route of decision-making steps, where sequential questions are made,

and routes are followed depending on the answer. The number of questions asked before reaching a predictive outcome, or the depth of the tree, can vary from one model to another. As depth increases, the model becomes more complex, with the risk of overfitting the model and obtaining unreliable results (Liberman, 2017).

A random forest is simply a combination of several decision trees than can be used to make classification predictions. As the article *What is Random Forest* from IBM (n.d.) explains, the random forest model is an ensemble of decision trees that uses bagging, which consists of taking with replacement different random samples of data from the training set, training independent models and generating a majority prediction that reduces variance. A key difference between random forest and standalone decision trees IBM (n.d.) identifies is that random forests operate with a subset of features, while decision trees consider all features.

The reason why we chose to use a random forest model instead of simply a decision tree lies in the fact that random forest can help to reduce overfitting without a significant increase in errors. This is because the model combines different decision trees that each include a different number of variables, thus reducing variance (Liberman, 2017).

In order to develop a decision tree, three parameters need to be chosen: number of features, number of decision trees and node size (IBM, n.d.). In our case we have chosen to generate 500 trees, have a minimum node size of 20 (minimum number of observations to be included in each node before doing a further partition), and use a Gini split for each decision tree, to try between 2 and 8 features to be included in the decision trees in order to develop the optimal model.

4) Ensemble of Logit, KNN and Decision Trees

The final models we have chosen to include in our study were two ensembles of Logit, KNN and decision trees, our three machine learning models we have used individually. The reason to do so was that ensembles are generally known for making more accurate predictions and having a higher performance. Since we have already tested each of the models individually, we wanted to also include in our study a model that combined all three techniques to analyse their predictive power together. An ensemble combines the individual outputs from the models include to generate a single output, with the objective to make better predictions. Although several approaches can be taken with ensembles, we decided to carry out stacking for each of the three included models, although results were only able to be obtained using decision trees and logit as stacking models, due to high computational capacity required with KNN.

Kundu (2022b) in its article *The Complete Guide to Ensemble Learning* mentions that the key advantage of ensemble models is the fact that prediction error variance is reduced substantially given that combined results are more strong than standalone outputs. An ensemble combines the

different information provided from each model, therefore diversity between them is needed. As we will see in the results section, prediction outputs for each of the individual models we have used are diverse, meaning that an ensemble formed by them complements well. Kundu (2022b) also explains that ensemble models are a good approach when not enough data is available. As our dataset was limited due to the shortage of publicly available information on Energy and Utilities sector companies, an ensemble is very appropriate for our study.

Stacking was the approach taken to ensemble our three models. In this approach, different subsamples of data are generated with replacement, called “bootstrap data”. These subsamples are taken as different datasets, and predictions are calculated over each of them. These prediction outputs are then converted into inputs of a chosen model called the “meta-classifier” which gives the final prediction correcting the behaviour of the previous predictions (Kundu, 2022b). In our study, we decided to create initially three ensembles using stacking, one for each of the individual models chosen, and that the search for hyperparameters of each of this model was done randomly. At the end, stacking was only able to be performed for stacking with logit and decision trees, due to computational errors as mentioned previously.

4. RESULTS

In this section of our paper, we are going to explain the results we have obtained from our study using R studio and programming in R, divided in two sections according to our two objectives.

4.1 EXPLANATORY MODEL

In the explanatory part of our analysis, we first compared what logit regression model was the best in explaining the outcome for each of the observations, to make sure the results where as accurate as possible. We developed three models:

(1) Including all 8 raw financial variables
(2) Instead of using the 8 raw variables, we removed EV on a standalone basis and substituted EBITDA, EBIT, and Revenue, by the corresponding ratios EV/EBITDA, EV/EBIT and EV/Revenue. This was because we wanted to test whether ratios improved the performance of the explanatory model and we wanted to introduce some interaction effects into the model.
(3) We removed from our financial variables Revenue and EBITDA, and left EBIT as a measure of profitability only. The reason we did this was that we felt that Revenue, EBITDA and EBIT are highly linked in financial statements, so a possibility existed of them being too highly correlated. We chose to use EBIT and not Revenue because the latter does not consider costs. On the other hand, we chose EBIT over EBITDA because we are working with Energy and Utilities companies, which are characterised by having high capital expenditures. Since depreciation can be used as a proxy for capital expenditure, EBIT was considering this important expense, but EBITDA was not.

To compare between the three models, the Akaike Information Criterion (AIC) metric was used. Bevans (2022) explains that AIC is a frequently used metric in statistics in order to select the model that shows the best results out of different combinations of variables. Lower values of AIC demonstrate that a model is a better fit for the data used, since its value increases as unnecessary variables are included. The formula for AIC is the following (Bevans, 2022):

$$AIC = 2K - 2Ln(L) \quad \textbf{(3)}$$

Where K is the number of independent variables (in our case 8) and L quantifies the likelihood of the model obtaining the correct y value observed.

The outcome of AICs of our 3 models chosen where the following:

(1) 1187.026
(2) 1247.036

(3) 1228.248

Based on this criterion, the explanatory model chosen to determine the significance of the different financial variables on our target variable and its marginal effects was the full model that included all variables and that had no interaction effects (model 1). This demonstrated that the model was not penalised by including all 8 financial variables, demonstrating that all are necessary in explaining the outcome of company being target or not in the Energy and Utilities Sector, and that including all variables make the model of a better fit.

Table 3. Summary for Chosen Explanatory model

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.7405     0.2001 -13.694  < 2e-16 ***
data$`Enterprise Value`       0.6592     0.5948   1.108  0.26773
data$Revenue                  1.3014     0.6152   2.115  0.03440 *
data$EBIT                   -17.9366     3.0142  -5.951 2.67e-09 ***
data$EBITDA                  16.1768     2.7389   5.906 3.50e-09 ***
data$`Total Assets`          -5.1373     1.7171  -2.992  0.00277 **
data$`Long Term Debt`         1.0367     0.6375   1.626  0.10393
data$`Cash & ST Investments` -1.8849     0.8486  -2.221  0.02635 *
data$`Price to Earnings`     -1.2371     0.4799  -2.578  0.00994 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Source: Own elaboration with R from FactSet Data

Table 3 was obtained from R studio and shows the different beta values calculated for our logit explanatory model. As explained above, this cannot be taken as the marginal effects due to the nature of the model. Instead, the valuable information from the table is the significance of each of the financial variables. We can divide the most significant variables in explaining whether a company in the Energy and Utilities sector will become a target or not in four Tiers, being Tier 1 the most significant, as shown in **Table 4.**

Table 4. Tiers of variable significance in our explanatory model

Tier 1	EBIT, EBITDA
Tier 2	Total Assets, Price to Earnings
Tier 3	Revenue, Cash and ST Investments
Tier 4	EV, Long Term Debt

Source: Own elaboration

The results obtained show the profitability (measured by EBIT and EBITDA) is the most important determinant on whether a company will become target or not. Second, Energy and Utilities sector public companies are also affected by the amount of total assets in its balance sheet (therefore capital expenditure) and by its trading value based on P/E. Revenue as a measure of amount of sales and cash as a measure of liquidity have less importance on the determination of

a target. Finally, the size of a company measured by EV and the riskiness of a company measured by long term debt are not significant on the results of a company being target.

The direction in which each of the variables affects whether a company is target or not can be seen with the marginal effect calculation on the average values of the different explanatory variables, as shown in **Table 5.** A positive value indicates that a unitary increase in one of the explanatory variables, while the rest remain constant, will increase the probability of a company being selected as target. The opposite occurs with a negative value, where a unitary increase will reduce the probability of a company being a target in an M&A transaction.

Table 5. Marginal effects for chosen Explanatory model

```
Marginal Effects:
                                    dF/dx
data$`Enterprise Value`          0.037541
data$Revenue                     0.074117
data$EBIT                       -1.021500
data$EBITDA                      0.921280
data$`Total Assets`             -0.292572
data$`Long Term Debt`            0.059040
data$`Cash & ST Investments` -0.107345
data$`Price to Earnings`        -0.070451
```

Source: Own elaboration with R from FactSet Data

From the marginal effects we can conclude from each of the 8 financial variables the results presented in **Table 6.** Commenting on them, we can see that public companies in the Energy and Utilities sector will more likely be targeted in M&A transactions if they have a higher size and higher levels of sales. Furthermore, profitability is variable. While companies with higher EBITDAs are preferred as targets, they are preferred with lower EBITs. The reasoning behind this could be regarding depreciation. Lower EBITs may be preferred because acquirers prefer companies with higher depreciations. Also, the fact that a company with a larger amount of assets in its balance sheet is less likely to become a target could be interpreted as acquirers preferring companies that have obsolete assets in their balance sheets, that are decreasing in value due to depreciation. This makes sense as acquirers may want to buy this type of companies in order to invest in them. Regarding risk and liquidity measured by debt and cash respectively, companies are more likely to be a target if they have higher levels of debt and are less liquid. We find this result interesting, as this type of companies tend to imply higher risk for acquirers. We can therefore conclude from here that acquirers may be focusing on other capabilities, such as the ability to generate synergies, and not pay that much attention to the riskiness of the companies, given that our dataset are public companies and therefore are considered established in the market. Another reason could also be that having more debt means more risk and therefore a lower price may be able to be paid, being this a priority for acquirers. Furthermore, both debt and cash are not very significant in the determination of the explanatory variable so their power on the decision is

not strong. Finally, a company is preferred as target if it has a lower P/E ratio, which makes sense as this means that the company is more likely undervalued compared to the market.

Table 6. Effect on Target Variable of Explanatory variables

Explanatory Variable	**Effect on Target Variable**
Enterprise Value	A larger EV implies *higher* likelihood of being a target
Revenue	Higher revenue implies *higher* likelihood of being a target
EBIT	Higher EBIT implies a *lower* likelihood of being a target
EBITDA	Higher EBITDA implies a *higher* likelihood of being a target
Total Assets	Higher assets imply a *lower* likelihood of being a target
Long Term Debt	Higher levels of debt imply a *higher* likelihood of being a target
Cash and ST Investments	Higher levels of cash imply a *lower* likelihood of being a target
Price to Earnings	Higher P/E imply a *lower* likelihood of being a target

Source: Own elaboration based on Explanatory model results

To conclude this section to answer our first objective and considering both the significance of the variables and the marginal effects, we can summarize that what most influences the decision in our set of companies of being a target in an M&A transaction is the level of profitability. While acquirers prefer companies with higher EBITDAs, they prefer companies with lower EBITs, implying that the difference lies beneath depreciation, and this could be related to the fact of having assets that are becoming obsolete and therefore need the investment of the acquirer to keep operating.

4.2 PREDICTIVE MODELS

This section gives the performance results of each of the five different models tested with our dataset, aiming to answer our second objective of identifying which model is the most accurate in making predictions of M&A targets across public companies in the Energy & Utilities sector. The section starts with a presentation of the individual results for each of the models selected and finalises with a comparison between models to select the best one.

4.2.1 LOGIT

Our first model chosen as a predictor is the logit model. After dividing our dataset into our training and test sets (70% and 30% respectively) and choosing 0,7 as our cut off as explained in our methodology section, we obtained performance results for our logit model including all 8 financial variables in our study.

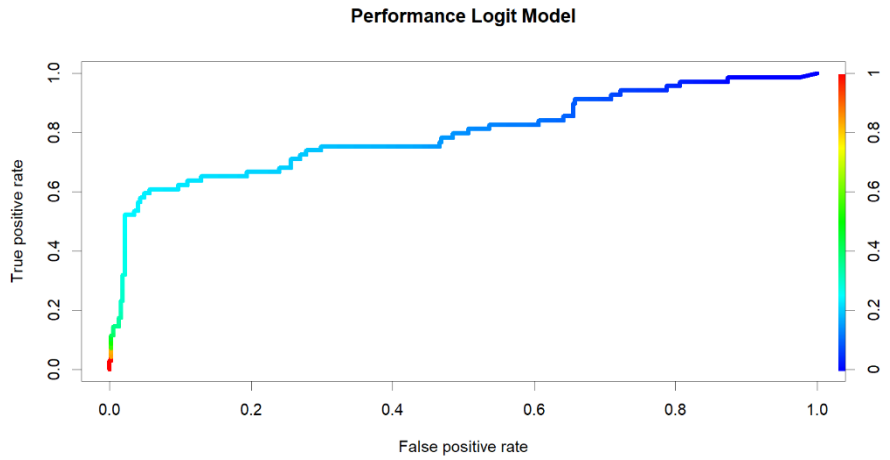
Table 7. Confusion matrix and statistics for Logit Predictive model

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 370  65
         1   1   4

               Accuracy : 0.85
                 95% CI : (0.8132, 0.8821)
    No Information Rate : 0.8432
    P-Value [Acc > NIR] : 0.3767

                  Kappa : 0.0888

 Mcnemar's Test P-Value : 8.851e-15

            Sensitivity : 0.057971
            Specificity : 0.997305
         Pos Pred Value : 0.800000
         Neg Pred Value : 0.850575
             Prevalence : 0.156818
         Detection Rate : 0.009091
   Detection Prevalence : 0.011364
      Balanced Accuracy : 0.527638

       'Positive' Class : 1
```

Source: Own elaboration with R from FactSet Data

From the confusion matrix we can first see that the Accuracy of the model is good, being 0,85 and slighlty above the no information rate (0,8432), which is used as a benchmark to compare the specific accuracy of our model with what would be obtained based if all observations where classified as the majority case ("non-target" in our case). While our classification error is of 15%, our type I error (or the proportion of false positives) is very low, of only 0,27%. While specificity here is very high (0,9973), sensitivity is very low (0,05791). Because of this, this model has a very low risk of identifying "non-target" companies as "target" companies, therefore there is a very low probability for investors to make the wrong investment. However, it is highly likely that it will miss "target" opportunities and therefore investors loose the chance to generate profits.

Figure 8. ROC Curve for Logit Predictive model



Source: Own elaboration with R from FactSet Data

Figure 8 shows the ROC curve for our logit model. Visually, we can see that the model seems quite accurate, since its line is plotted towards the left corner of the chart. The colour code on the left symbolizes the cut-off or threshold chosen. As the cut off value decreases, the proportion of false positives increases, leading to a higher Type I error. Finally, the AUC metric for this model is of 0.794992, which is very acceptable.

4.2.2 KNN

Our second model analysed is KNN, for which we use a cross validation in order to obtain the optimum number of K neighbours the model will use to assign a class to our test observations. After carrying out 3 repeats to ensure the results are accurate, the optimum k is of 33 data points. We then use ROC to choose the optimal model.

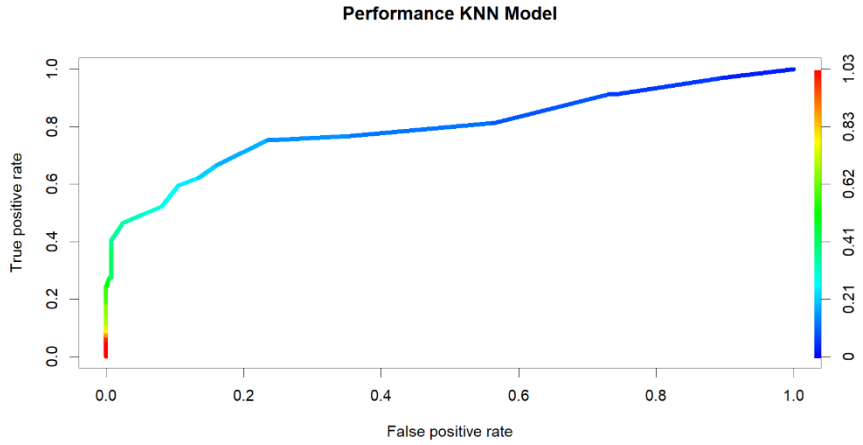
Table 8. Confusion matrix and Statistics for KNN predictive model

```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 370  52
         Y   1  17

               Accuracy : 0.8795
                 95% CI : (0.8454, 0.9085)
    No Information Rate : 0.8432
    P-Value [Acc > NIR] : 0.01852

                  Kappa : 0.3485

 Mcnemar's Test P-Value : 6.51e-12

            Sensitivity : 0.24638
            Specificity : 0.99730
         Pos Pred Value : 0.94444
         Neg Pred Value : 0.87678
             Prevalence : 0.15682
         Detection Rate : 0.03864
   Detection Prevalence : 0.04091
      Balanced Accuracy : 0.62184

       'Positive' Class : Y
```

Source: Own elaboration with R from FactSet Data

The confusion matrix in this case has an accuracy of 0,8795, which is considered quite high, and quite above the no information rate value of 0,8432. Again, the type I error which shows the proportion of false positives is minimal, of 0,27%, and the classification error of 12,05%. Specificity is very close to 1, reducing the risk to almost 0 of investors putting money on the wrong companies. On the other hand, sensititivity is quite acceptable, as almost a quarter of companies that are "targets" will be identified correctly and investors will have the opportunity to invest in them.

Figure 9. ROC Curve for KNN Predictive model



Source: Own elaboration with R from FactSet Data

In **Figure 9** we can see the plot of the ROC Curve for the KNN predictive model. The models plot appears quite accurate, and has an AUC value of 0,7961639, which is fairly good.

4.2.3 RANDOM FOREST

The third model chosen is an ensemble of several decision trees, or random forest, for which we have chosen to generate 500 trees, with a minimum node size of 20 and with the optimum feature selection chosen between 2 and 8 features. After repeating the process 3 times to ensure accurate results, the optimal model obtained using ROC criteria has been one with an mtry value of 4, which means 4 features.

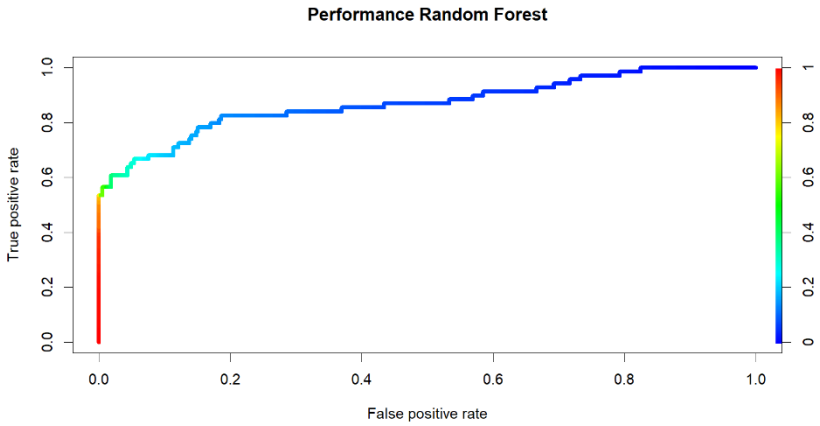
Table 9. Confusion matrix and statistics for Random Forest Predictive model

```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 367  30
         Y   4  39

               Accuracy : 0.9227
                 95% CI : (0.8937, 0.9459)
    No Information Rate : 0.8432
    P-Value [Acc > NIR] : 4.540e-07

                  Kappa : 0.6549

 Mcnemar's Test P-Value : 1.807e-05

            Sensitivity : 0.56522
            Specificity : 0.98922
         Pos Pred Value : 0.90698
         Neg Pred Value : 0.92443
             Prevalence : 0.15682
         Detection Rate : 0.08864
   Detection Prevalence : 0.09773
      Balanced Accuracy : 0.77722

       'Positive' Class : Y
```

Source: Own elaboration with R from FactSet Data

As seen in **Table 9,** the confusion matrix shows a very high accuracy of 0,9227, substantially high over the no information rate of 0,8432, which demonstrates the model has very good performance. Classification error is very low, of only 7,73% and the error of false positives is of 1,08%. Specificity is substantially high (0,98922) and sensititivity has a very acceptable value of 0,56522. Because of this, there is very low risk of investment in false positive companies and there are more than 50% of companies identified correctly as “target”, implying high opportunities for investors to make a profit.

Figure 10. ROC Curve for Random Forest Predictive Model



Source: Own elaboration with R from FactSet Data

Figure 10 clearly shows the high predictive power of the random forest model due to its shape moving towards the top left side of the chart. Its AUC value is of 0,8696043 which is very high.

4.2.4 ENSEMBLE MODELS

The last models analysed were two ensembles of logistic regression, decision trees and KNN. We used again cross validation to generate the optimal model, repeating the process 3 times and with a random search of hyperparameters, and using once again the 8 financial variables as independent variables. The difference with the other models analysed was that the criteria used to select the best models was accuracy instead of ROC or AUC. This was because this type of models do not allow to calculate ROC or AUC because of the nature of the algorithm.

Two ensembles were done with stacking. The stacking was wanted to be performed with each of the three models incorporated in order to be able to compare which one was the best. However, due to computational efforts required, results for KNN as the stacking algorithm were not obtained.

Table 10 shows the results for the ensemble using logit regression as the stacking model. As seen, the accuracy is very high of 0,9227, and well above the no information rate (0,8432). Specificity is of almost 1, again reducing the error of false positive to the minimum, and sensitivity is very acceptable (0,52174), giving the possibility to investors to possibly know half of which companies will become targets. While the classification error is low, 7,73%, the type I error is even lower (0,27%).

Table 10. Confusion Matrix and Statistics for Ensemble with Logit stacking Predictive model

```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 370  33
         Y   1  36

               Accuracy : 0.9227
                 95% CI : (0.8937, 0.9459)
    No Information Rate : 0.8432
    P-Value [Acc > NIR] : 4.540e-07

                  Kappa : 0.6398

Mcnemar's Test P-Value : 1.058e-07

            Sensitivity : 0.52174
            Specificity : 0.99730
         Pos Pred Value : 0.97297
         Neg Pred Value : 0.91811
             Prevalence : 0.15682
         Detection Rate : 0.08182
   Detection Prevalence : 0.08409
      Balanced Accuracy : 0.75952

       'Positive' Class : Y
```

Source: Own elaboration with R from FactSet Data

Table 11 includes the results obtained from the second stacking model considered, using decision trees as the meta-classifier. Here, the accuracy is slighlty lower, of 0,9045, but still well above the no iformation rate. Specificity is of 0,97035 and sensitivity of 0,55072, again accetable values for our models. Finally, classification error is of 9,55% and error type I of 2,97%.

Table 11. Confusion matrix and statistics for Ensemble with decision trees stacking Predictive model

```
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N 360  31
         Y  11  38

               Accuracy : 0.9045
                 95% CI : (0.8732, 0.9303)
    No Information Rate : 0.8432
    P-Value [Acc > NIR] : 0.0001171

                  Kappa : 0.5908

Mcnemar's Test P-Value : 0.0033704

            Sensitivity : 0.55072
            Specificity : 0.97035
         Pos Pred Value : 0.77551
         Neg Pred Value : 0.92072
             Prevalence : 0.15682
         Detection Rate : 0.08636
   Detection Prevalence : 0.11136
      Balanced Accuracy : 0.76054

       'Positive' Class : Y
```

Source: Own elaboration with R from FactSet Data

4.2.5 COMPARISON BETWEEN MODELS

In order to conclude and answer our second objective of which model makes the best prediction of M&A targets in the Energy and Utilities sectors, we have decided to synthesize the results for comparison as shown in **Table 12.** Colour coding shows how good the results are, being a darker green a better performance while a white colour a less precise model. Given this we can see that overall, the logit stacking ensemble was the most accurate model, with the highest accuracy, lowest classification and type I errors, highest specificity, and almost highest sensitivity (only a few points below random forest and decision tree stacking ensemble). However, we were not able to calculate the AUC of this model. Because of this, if we had to choose the model in terms of AUC, we would choose random forest, due to its much higher value above logit and KNN.

The least accurate model we could say would be logit, with the highest classification error and lowest accuracy and AUC. These results make sense as even though logit is a widely used model for classification problems, its performance is quite limited.

Table 12. Summary Comparison between predictive models

Model	Accuracy	Classification Error	Type I error	Specificity	Sensitivity	AUC
Logit	0,85	15%	0,27%	0,997305	0,057971	0,794992
KNN	0,8795	12,05%	0,27%	0,9973	0,24638	0,7961639
Random Forest	0,9227	7,73%	1,08%	0,98922	0,56522	0,8696043
Logit Stacking Ensemble	0,9227	7,73%	0,27%	0,9973	0,52174	-
Decision Tree Stacking Ensemble	0,9045	9,55%	2,97%	0,97035	0,55072	-

Source: Own Elaboration using FactSet Data

5. CONCLUSIONS

After carrying out our explanatory and predictive analysis to answer our two objectives, the answers obtained have been the following. We can first say that profitability is the financial factor that most affects the determination of M&A targets in the Energy and Utilities sectors. Acquirers prefer companies with higher EBITDAs but lower EBITs, and this difference in preference lies within depreciation. Acquirers prefer companies with higher depreciation, and therefore that need investment in assets. For our second objective, we can conclude that more complex models tend to give better prediction performance. While a logit stacking ensemble proves to be the best performing model above all tested, this has the limitation of not being able to obtain its AUC. If we were to base our best models in terms of this metric, then random forest would be the most precise.

While the results concluded from the study give a fair insight to investors on what financial variables affect the selection of M&A targets in the Energy and Utilities Sector, and suggestions on what model should be used by investors to predict targets, M&A target selection is a complex field affected by infinite variables. While financial variables can be a key part of the determination of targets, there are many other variables that affect target selection. This includes macroeconomic factors, stage of the business cycle, possibility to create synergies, previous ownership of companies, qualitative factors such as climate risk exposure, among others. Also, it includes industry specific variables like the price of oil or electricity, or the amount of oil production in the case of Energy and Utilities companies. Because of this, the results of this study are not determinant, but complementary to other analysis that can be performed to select M&A targets in the Energy and Utilities sector. Predicting if a company will become target or not requires not only tools to assist on the prediction, but also a lot of experience in the field.

Even though the study carried out has been done with due diligence and as precise as possible, it is true to say that the study has had some limitations. The fact that we centred our study in the Energy and Utilities sector made it more precise than using several industries, as we reduced the effect of industry having an impact on target selection. However, the availability of data was very limited, and this led to some impreciseness. Most M&A targets tend to be private companies due to size, but there is a lack of public information on these types of companies. Because of this, we were forced to work only with public companies, which were more limited in number. Also, we had no choice but to use a time frame of 10 years, which was acceptable but meant that target selection could be affected by business cycles in the past. Having a shorter, more recent data frame could have provided better results. Furthermore, having a limited dataset did not allow us

to centre our study in a specific country or area. Because of this, macroeconomic factors in the continents may have impacted our results.

Finally, there is room for continuation for this study. First, more industry specific metrics related to the Energy and Utilities sector could be used as variables that explain or predict target companies. Second, a higher number of models could have been used for a larger comparison. To end, if more data is accessible then the study could become more specific and precise, by centring the dataset on a specific continent.

6. BIBLIOGRAPHY

Andriuškevičius, K., & Štreimikienė, D. (2021). Developments and trends of mergers and acquisitions in the energy industry. *Energies*, *14*(8), 2158.

Aramyan, H. (2021, September 21) *Predicting M&A Targets Using Machine Learning Techniques.* Refinitiv. Retrieved April 26, 2023 from https://developers.refinitiv.com/en/article-catalog/article/prediction-of-m-and-a-targets-to-generate-portfolio-returns#Conclusion

Bain & Company (2023). *Global M&A Report 2023.* https://www.bain.com/globalassets/noindex/2023/bain_report_global_m_and_a_report_2023.pdf

Bevans, R. (2022, November 18). *Akaike Information Criterion | When & How to Use It (Example)*. Scribbr. Retrieved May 18, 2023, from https://www.scribbr.com/statistics/akaike-information-criterion/

CFA Institute (2022). *Mergers and Acquisitions.* https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/mergers-acquisitions#:~:text=An%20acquisition%20is%20the%20purchase,by%20the%20form%20of%20integration.

Deloitte (2023). *Oil and Gas M&A Outlook 2023: Pivoting for Change.* https://www2.deloitte.com/us/en/pages/energy-and-resources/articles/oil-and-gas-mergers-and-acquisitions.html

DePamphilis, D. (2003). Acquisitions and Other Restructuring Activities: An Integrated Approach to Process, Tools, Cases, and Solutions. *Butterworth-Heinemann.*

Feldman, E. R., & Hernandez, E. (2022). Synergy in Mergers and Acquisitions: Typology, Life Cycles, and Value. *Academy of Management Review*, *47*(4), 549-578.

Furenmo, G. (2020). Predicting Corporate Takeover Outcomes Using Machine Learning. *Bachelor's Thesis in Economics from Lund University.*

IBM (n.d.). *What is Random Forest.* Retrieved May 21, 2023 from https://www.ibm.com/topics/random-forest

IMAA Institute (n.d.). *M&A Statistics by industries.* Retrieved April 9, 2023 from https://imaa-institute.org/mergers-and-acquisitions-statistics/ma-statistics-by-industries/

Kapil, S., & Dhingra, K. (2021). Understanding Determinants Of Domestic Mergers And Acquisitions Through Literature Review. *Indian Journal of Finance and Banking*, *6*(1), 31-57.

Kim, W. G., & Arbel, A. (1998). Predicting merger targets of hospitality firms (a Logit model). *International Journal of Hospitality Management*, *17*(3), 303-318.

Kolostyak, S. (2021, December 2). *What Does M&A Mean for Investors?* Morningstar. Retrieved April 23, 2023 from https://www.morningstar.co.uk/uk/news/217257/what-does-ma-mean-for-investors.aspx

KPMG (2007). *The Determinants of M&A Success.*

KPMG (2021). 2021 was a blowout year from global M&A. https://advisory.kpmg.us/articles/2021/blowout-year-global-ma.html

Kundu, R. (2022a, September 13). *Confusion Matrix: How To Use It & Interpret Results [Examples]*. Retrieved May 20, 2023 from https://www.v7labs.com/blog/confusion-matrix-guide

Kundu, R. (2022b, March 1) *The Complete Guide to Ensemble Learning.* Retrieved May 21 from https://www.v7labs.com/blog/ensemble-learning

Leepsa, N. M., & Mishra, C. S. (2016). Theory and practice of mergers and acquisitions: Empirical evidence from Indian cases. *IIMS Journal of management science*, *7*(2), 179-194.

Liberman, N. (2017, Jan 27). *Decision Trees and Random Forests.* Towards Data Science. Retrieved May 21, 2023 from https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

Malik, M. F., Anuar, M. A., Khan, S., & Khan, F. (2014). Mergers and acquisitions: A conceptual review. *International Journal of Accounting and Financial Reporting*, 4(2), 520.

McKinsey & Company. (2010). *Opening the aperture 1: A McKinsey perspective on value creation and synergies.* https://www.mckinsey.com/client_service/organization/latest_thinking/~/media/D74F0B9DCDAB4EAE918D2D578E7C7AF7.ashx

McKinsey & Company (2022, December 9). *Ready, set, grow: Winning the M&A race for renewable developers.* Retrieved April 9, 2023 from https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/ready-set-grow-winning-the-m-and-a-race-for-renewables-developers#/

MET Group (2022, February 28). *Energy Sector Definition: How does the energy industry work?*. Retrieved April 8, 2023 from https://group.met.com/en/media/energy-insight/energy-sector-industry

MSCI & S&P Global (2023). *Global Industry Classification Sector* https://www.msci.com/documents/1296102/11185224/GICS+Sector+Definitions+2023.pdf/822305c6-f821-3d65-1984-6615ded81473?t=1679088764288

M&A Research Centre at Cass Business School, City University London & Intralinks (2016). *Attractive M&A Target: Part 1 What do buyers look for?*. Retrieved April 9, 2023 from https://www.mergermarket.com/assets/Attractive_M&A_Targets_PART%201_v2.pdf

Narkhede, S. (2018, June 26). *Understanding AUC – ROC Curve.* Towards Data Science. Retrieved May 20, 2023 from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5#:~:text=the%20multiclass%20model%3F-,What%20is%20the%20AUC%20%2D%20ROC%20Curve%3F,capable%20of%20distinguishing%20between%20classes.

Nicholas Center for Corporate Finance & Investment Banking (2019, December 15). From ML to M&A. Ten M&A Target Predictions through a Machine Learning Model. *Wisconsin School of Business.*

Osborne, S., Katselas, D., & Chapple, L. (2012). The preferences of private equity investors in selecting target acquisitions: An international investigation. *Australian Journal of Management*, *37*(3), 361-389.

Peat, M., & Stevenson, M. (2008). Predicting Australian takeover targets: a logit analysis. *The 6th INFINITI Conference on International Finance*.

Raschka, S. (2018). STAT 479: Machine Learning Lecture Notes. *University of Wisconsin-Madison. Department of Statistics.*

Roberts, A., Wallace, W., & Moles, P. (2003). Mergers and acquisitions. *Pearson Education.*

Sebastiano, M. (2021). *Mergers and Acquisitions: expected vs actual performance. A set of case study assessments* (Doctoral dissertation, Politecnico di Torino).

Skevofylakas, M. (2022, July 12). *Modelling and Evaluation – Artificial Intelligence regression and classification performance metrics.* Refinitiv. Retrieved May 20, 2023 from https://developers.refinitiv.com/en/article-catalog/article/modelling-and-evaluation-artificial-intelligence-regression-and-classification-performance-metrics

Stock, J.H., & Watson, M.W. (2015). *Introduction to Econometrics.* Pearson.

S&P Global (2020, February 24). *What is Energy Transition?*. Retrieved April 8, 2023 from https://www.spglobal.com/en/research-insights/articles/what-is-energy-transition

Tamosiuniene, R., & Duksaite, E. (2009). The importance of Mergers and Acquisitions in Today's Economy. *KSI Transactions on Knowledge Society*, *2*(4), 11-15.

Wilson, T. (2022, July 27). *Oil and gas majors: time for a transformative clean energy deal?* Financial Times. Retrieved April 9, 2023 from https://www.ft.com/content/760ba17c-3437-45eb-841a-e5980bf3ae22

7. ANNEX: CODE

```
########TFG ANALYTICS


#Read Dataset and Fix Directory

library(readxl)

data<- read_excel("C:/Users/lucia/Desktop/E2-ANALYTICS/TFGS/ANALYTICS/Dataset TFG Analytics.xlsx")

View(data)


#Load Libraries

library(ROCR)

library(pROC)

library(dplyr)

library(RColorBrewer)

library(ggplot2)

library(GGally)

library(Hmisc)

library(corrplot)

library(tidyr)

library(tidyverse)

library(caret)

library(e1071)

library(rpart)

library(caretEnsemble)

library(mfx)

library(plotROC)


##Summary of data

summary(data)

#observe very disperse values in size, we will need to standardize them when doing our explanatory and predictive study, in order to carry out our model in an optimal way
```

```
#Identify any missing values
sapply(data,anyNA)
#we see there are no missing values

####DATA ANALYSIS
##CATEGORIC ANALYSIS

#chart to count number of businesses per industry
ggplot(data, aes(x =`FactSet Industry`)) +
  geom_bar(color = 'darkblue', fill = 'darkblue') +
  labs(x = "Industry", y = "Count") + theme_minimal()

Industrycount <-table(data$`FactSet Industry`)
Industrytable <- sort(Industrycount, decreasing = TRUE)
print(Industrytable)
view(Industrytable)
num_distinctindustries <- length(unique(data$`FactSet Industry`))
print(num_distinctindustries)

Countrycount <-table(data$Country)
Countrytable <- sort(Countrycount, decreasing = TRUE)
print(Countrytable)
view(Countrytable)
num_distinctcountries <- length(unique(data$Country))
print(num_distinctcountries)


###NUMERIC ANALYSIS
#cORRELATION MATRIX OF NUMERIC VARIABLES
#select numeric variables (target at the moment is still numeric for our analysis, we will then change it to factor)
numericdata<-data
```

```
####numericdata$`Target (Yes/No)`<-as.numeric(numericdata$`Target (Yes/No)`)
####view(numericdata)
numericdata<-select_if(data, is.numeric)
view(numericdata)

col<- colorRampPalette(c("darkblue", "white", "lightblue"))(10)
corrmatrix <-cor(numericdata)
#see correlation values
round(corrmatrix,2)
#visualiZe correlation
corrplot(corrmatrix, type="upper", order="hclust",
         col=col, tl.col="darkblue")

##observe none of the variables are correlated with our target variable, we don't delete any
##some variables are correlated between each other, this is normal as they are dependent
##we don't eliminate any variable because of the nature of our study

# BOXPLOT BY INDUSTRY
par(mar = c(8,4, 0.1, 0.1))
boxplot(numericdata$`Enterprise Value` ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "EV",
        col = "lightblue",
        las=2,
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$Revenue ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "Revenue",
        col = "lightblue",
        las=2,
```

```
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$EBIT ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "EBIT",
        col = "lightblue",
        las=2,
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$EBITDA ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "EBITDA",
        col = "lightblue",
        las=2,
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$`Total Assets`~ data$`FactSet Industry`,
        xlab = "",
        ylab = "Total Assets",
        col = "lightblue",
        las=2,
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$`Long Term Debt` ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "Long Term Debt",
        col = "lightblue",
        las=2,
```

```
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$`Cash & ST Investments` ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "Cash & ST Investments",
        col = "lightblue",
        las=2,
        cex.axis=0.6,
        cex.lab=0.6)

boxplot(numericdata$`Price to Earnings` ~ data$`FactSet Industry`,
        xlab = "",
        ylab = "P/E",
        col = "lightblue",
        las=2,
        cex.axis=0.6,
        cex.lab=0.6)

###Once data analysis is done, we scale variables and turn target into factor

#scale data
data$`Enterprise Value`<-scale(data$`Enterprise Value`)
data$Revenue<-scale(data$Revenue)
data$EBIT<-scale(data$EBIT)
data$EBITDA<-scale(data$EBITDA)
data$`Total Assets`<-scale(data$`Total Assets`)
data$`Long Term Debt`<-scale(data$`Long Term Debt`)
data$`Cash & ST Investments`<-scale(data$`Cash & ST Investments`)
data$`Price to Earnings`<-scale(data$`Price to Earnings`)
summary(data)
```

```
data$`Target (Yes/No)`<-as.factor(data$`Target (Yes/No)`)

summary(data)

#####EXPLICATIVE LOGIT MODEL

fullmodel<-glm(`Target (Yes/No)`~data$`Enterprise Value`+data$Revenue+data$EBIT + data$EBITDA + data$`Total Assets` + data$`Long Term Debt` + data$`Cash & ST Investments` + data$`Price to Earnings`,data=data, family=binomial(logit))

fullmodel

summary(fullmodel)

AIC(fullmodel)

#1187.026


model1<-glm(`Target (Yes/No)`~data$`Enterprise Value`/data$EBITDA+data$`Enterprise Value`/data$Revenue+data$`Enterprise Value`/data$EBIT + data$`Total Assets` + data$`Long Term Debt` + data$`Cash & ST Investments` + data$`Price to Earnings`,data=data, family=binomial(logit))

model1

summary(model1)

AIC(model1)

#1247.036


model2<-glm(`Target (Yes/No)`~data$`Enterprise Value`+data$EBIT + data$`Total Assets` + data$`Long Term Debt` + data$`Cash & ST Investments` + data$`Price to Earnings`,data=data, family=binomial(logit))

model2

summary(model2)

AIC(model2)

#1228.248


###Conclude the full model is the best because of having a lower AIC


##marginal effect calculation to see how a change in each variable affects the probability of being a target for full model

mfx::logitmfx(formula = fullmodel, data = data)
```

```
##################PREDICTION MODELS

#DATA PARTITION

#Set seed and generate aleatory numbers

RNGkind("Super", "Inversion", "Rounding")

set.seed(123)

#Partition Training 70% y Test 30%

index<-createDataPartition(data$`Target (Yes/No)`, p=0.7, list=FALSE)

train<-data[index,]

test<-data[-index,]

summary(train)

summary(test)

#Fix cut-off value to establish thresholds for classifying targets and non-targets

cutoff<-0.7

####LOGIT MODEL

logitmodel<-glm(`Target (Yes/No)`~`Enterprise Value`+Revenue+EBIT+ EBITDA +`Total Assets` + `Long Term Debt` + `Cash & ST Investments` + `Price to Earnings`,data=train, family=binomial(logit))

test$predNuma<-predict(logitmodel, newdata=test, type="response")

test$predClasa<-ifelse(test$predNuma>=cutoff,1,0)

#Turn into factor to create confussion matrix

test$predClasa<-as.factor(test$predClasa)

test$`Target (Yes/No)`<-as.factor(test$`Target (Yes/No)`)

#Confussion Matrix

confusionMatrix(test$predClasa, test$`Target (Yes/No)`, positive="1")

#worse error is predicting that the company will be a target and then it is not --> investors could loose money
```

```
#false positive is the worst error
predlogit<-prediction(test$predNuma, test$`Target (Yes/No)`)

#ROC Curve
perflogit<-performance(predlogit,"tpr","fpr")
par(mar = c(5, 5, 5, 5))
plot(perflogit, colorize=TRUE,lwd = 5, main="Performance Logit Model")

#AUC
perflogit2<-performance(predlogit,measure="auc")
perflogit2@y.values
#AUC=0.794992

#####KNN

#Make 1 and 0 into Y and N respectively to make it more understandable
data$`Target (Yes/No)` = ifelse(data$`Target (Yes/No)`==1,"Y","N")
data$`Target (Yes/No)`<-factor(data$`Target (Yes/No)`)
summary(data$`Target (Yes/No)`)

##cross validation to obtain the optimum k number of nearest neighbors

repeats = 3
numbers = 10
tunel = 30

#PARTITION
RNGkind("Super", "Inversion", "Rounding")
set.seed(123)
index<-createDataPartition(data$`Target (Yes/No)`, p=0.7, list=FALSE)
train<-data[index,]
test<-data[-index,]
```

```
#CROSS-VALIDATION
RNGkind("Super", "Inversion", "Rounding")
set.seed(123)
x = trainControl(method = "repeatedcv", #cross validation for optimum K
                 number = numbers, #folds
                 repeats = repeats,       #repeats of cross validation
                 classProbs = TRUE,  #to obtain probabilities as an output
                 summaryFunction = twoClassSummary) #specify we are using a binary variable

knnfull<- train(`Target (Yes/No)`~`Enterprise Value`+Revenue+EBIT + EBITDA + `Total Assets` + `Long Term Debt` + `Cash & ST Investments` + `Price to Earnings`, data = train,
                method = "knn",
                preProcess = c("center","scale"),                    #standardize variables
                trControl = x,
                metric = "ROC",                  #use ROC to choose the best model
                tuneLength = tunel)

knnfull


#CONFUSION MATRIX
test_classknnfull<-predict(knnfull, newdata=test)
test_predknnfull<-predict(knnfull, newdata=test, type="prob")

confusionMatrix(data=test_classknnfull, test$`Target (Yes/No)`, positive="Y")

predfullknn_test<-prediction(test_predknnfull[,2],test$`Target (Yes/No)`,label.ordering = c("N","Y"))

#ROC CURVE
perf_testknnfull1<-performance(predfullknn_test, "tpr", "fpr")
```

```
plot(perf_testknnfull1, colorize=TRUE, main="Performance KNN Model", lwd = 5)

#AUC
perf_testknnfull2<-performance(predfullknn_test, "auc")
perf_testknnfull2@y.values
#auc=0.7961639



#####Random Forest
#already transformed target variable into factor Y and N
summary(data)

##Data partition
RNGkind("Super", "Inversion", "Rounding")
set.seed(123)

index<-createDataPartition(data$`Target (Yes/No)`, p=0.7, list=FALSE)
train<-data[index,]
test<-data[-index,]

#optimise mtry
# Setting up train controls, Gini index split and each node has at least 20 individuals
numbers <- 10
rep <- 3
tgrid <- expand.grid(
  .mtry = 2:8,
  .splitrule = "gini",
  .min.node.size =20)

RNGkind("Super", "Inversion", "Rounding")
set.seed(123)
```

```
x <- trainControl(method = "repeatedcv",                    # cross validation
                  number = numbers,       # number of folds for cross validation
                  repeats = rep,
                  verboseIter = TRUE,
                  classProbs = TRUE,            # classification probabilities
                  summaryFunction=twoClassSummary)



#model estimation
randomforestmodel <- train(`Target (Yes/No)`~`Enterprise Value`+Revenue+EBIT + EBITDA + `Total Assets` + `Long Term Debt` + `Cash & ST Investments` + `Price to Earnings`, data = train,
                  method = "ranger",                                #using the ranger algorithm that manages well large datasets
                  trControl = x,                                    #traincontrol
                  tuneGrid = tgrid,       # number of mtry values to try
                  num.trees=500,
                  metric="ROC" ,          # select the best model using ROC metric
                  importance="impurity"
)

randomforestmodel

#prediction
predclass<-predict(randomforestmodel, newdata=test, type="raw")

#confusion matrix
confusionMatrix(predclass, test$`Target (Yes/No)`, positive = "Y")

##ROC
predclass<-predict(randomforestmodel, newdata=test, type="prob")
predfullrandomforest<-prediction(predclass[,2],test$`Target (Yes/No)`,label.ordering = c("N","Y"))
```

```
perfrandomforest1<-performance(predfullrandomforest, "tpr", "fpr")
plot(perfrandomforest1, colorize=TRUE, main="Performance Random Forest", lwd=5)

#AUC
perfrandomforest2<-performance(predfullrandomforest, "auc")
perfrandomforest2@y.values
#AUC=0.8696043


######ENSEMBLE MODEL

#Random number generation
RNGkind("Super", "Inversion", "Rounding")
set.seed(123)

#remove spaces from variables to avoid errors
colnames(data) <- make.names(colnames(data))
view(data)

#DATA PARTITION
index = createDataPartition(data$Target..Yes.No., p = 0.7, list = F )
train = data[index,]
test = data[-index,]

#Create Supervisor Model
control <- trainControl(method="repeatedcv", number=10, repeats=3, classProbs=TRUE, search="random")

#Algorithm selection: decision tree, logit and knn
algorithmList <- c('rpart','glm','knn')

RNGkind("Super", "Inversion", "Rounding")
set.seed(123)
```

```
ensemblemodel <- caretList(Target..Yes.No.~Enterprise.Value+Revenue+EBIT+EBITDA+Total.Assets+Long.Term.Debt+Cash...ST.Investments+Price.to.Earnings, data=train, trControl=control, preProcess=c("center","scale"), methodList=algorithmList, tuneLength=10)


# stack using glm

stackControl <- trainControl(method="repeatedcv", number=10, repeats=3, savePredictions=TRUE, classProbs=TRUE, search="random")


RNGkind("Super", "Inversion", "Rounding")

set.seed(123)


stack.glm <- caretStack(ensemblemodel, method="glm", metric="Accuracy", trControl=stackControl)

print(stack.glm)


pred.glm<-predict(stack.glm, newdata=test)

confusionMatrix(pred.glm, test$Target..Yes.No., positive="Y")

#Accuracy=0.9227


# stack using decision trees

RNGkind("Super", "Inversion", "Rounding")

set.seed(123)


stack.rpart <- caretStack(ensemblemodel, method="rpart", metric="Accuracy", trControl=stackControl,tuneLength=5)

print(stack.rpart)


pred.rpart<-predict(stack.rpart, newdata=test)

confusionMatrix(pred.rpart, test$Target..Yes.No., positive = "Y")

#Accuracy=0.9045


#stack using knn
```

```
RNGkind("Super", "Inversion", "Rounding")

set.seed(123)


stack.knn <- caretStack(ensemblemodel, method="knn", metric="Accuracy", trControl=stackControl,tuneLength=5)

print(stack.knn)


pred.knn<-predict(stack.knn, newdata=test)

confusionMatrix(pred.knn, test$Target..Yes.No., positive = "Y")

###not able to perform, computational error
```