A Data-Driven Framework for Identifying Investment Opportunities in Private Equity

A Data-Driven Framework for Identifying Investment Opportunities in Private Equity

Samantha Petersone[1]∗, Alwin Tan[1]∗, Richard Allmendinger[1], Sujit Roy[1], James Hales[2]

[1] *The University of Manchester, Booth St W, Manchester, United Kingdom, M15 6PB ;*
[2] *NorthEdge Capital LLP, St Pauls House, 23 Park Square S, Leeds LS1 2ND*
∗ *Equal Contribution*
Email: richard.allmendinger@manchester.ac.uk

Abstract

The core activity of a Private Equity (PE) firm is to invest into companies in order to provide the investors with profit, usually within 4-7 years. To invest into a company or not is typically done manually by looking at various performance indicators of the company and then making a decision often based on instinct. This process is rather unmanageable given the large number of companies to potentially invest. Moreover, as more data about company performance indicators becomes available and the number of different indicators one may want to consider increases, manual crawling and assessment of investment opportunities becomes inefficient and ultimately impossible. To address these issues, this paper proposes a framework for automated data-driven screening of investment opportunities and thus the recommendation of businesses to invest in. The framework draws on data from several sources to assess the financial and managerial position of a company, and then uses an explainable artificial intelligence (XAI) engine to suggest investment recommendations. The robustness of the model is validated using different AI algorithms, class imbalance-handling methods, and features extracted from the available data sources.

Keywords: Explainable AI, Private Equity, Machine learning, Data Imbalance, Investment Leads Identification

1. Introduction

Private equity (PE) is a broad alternative investment class [1]. A PE investment usually involves the purchase of part or all of a company that is not publicly traded on a stock exchange, even though investors may occasionally involve in the privatisation of publicly listed company. PE investments deal with companies from all stages. The funds raised from the investment are often tied to some purpose ranging from the product development for companies in their early stages to expansion in operations for mature companies.

Over the past few decades, despite the rapid growth in the PE industry and the advancement in financial technology, studies focusing on the application of such techniques on investment decision support systems in the field of PE are still rare. This can be due to multiple reasons. Firstly, the **sparsity and scarcity of the historical deal data and financial data from small companies** has been considered as one of the major factors hampering advancements of data-driven methods in the PE market. According to the UK's audit exemption for small companies and micro entities[1], these companies are not legally obligated to disclose their financial statement in full. They have the option to submit an abridged ('simpler') account or choose not to disclose their financial statement to the public depending on the region. In addition, PE firms are also not required to disclose the deal investment and exit values publicly, further aggravating the issue of data sparsity and scarcity. This limits further exploration using data-heavy Artificial Intelligence (AI) method, such as deep learning [2].

Next, the degree of complexity involved in applying computational methods in PE is exacerbated by the **qualitative and intuitive nature of investment decision making**. PE managers often report their decisions being based more on their 'gut feeling' than hard-coded rules and data. An analysis has also been conducted on how the investment works based on entrepreneur profile or the idea [3]. The inability to quantify gut-feeling can also be associated with the important role played by the quality of the target companies management team and its relationships with the PE firm's partners. Naturally, such relations are hard to quantify, contributing to the slow progress of AI adaptation in the PE industry. Moreover, black-box AI methods are of limited use in the PE domain because of the high risks involved in investments (primarily boiling down to the risk of losing a significant amount of money). This urges the need for explainable AI (XAI) methods [4].

Apart from this, the **holding period for PE investment is relatively long**, spanning a few to many years. Therefore, the estimation for any models using the exit outcome, for example, return on investment specific to the type of exit, may be less relevant as the market environment changes at a rapid pace. As a result, the model is less meaningful as they may not accurately address the current investment environment.

Finally, **data imbalance** has been repeatedly reported as an issue by many academic literature across different types of studies related to PE, as the number of no-deal companies significantly exceeds the number of deal companies. This imbalance in data create further challenges in making biased prediction. For example, when models are trained to predict investment into a company, simply predicting no investment will be the best decision in most cases. Some studies have also used exit type data, which itself suffers from the survivorship bias and the under-representation of businesses going public after the deal compared to those who chose to stay private. Similarly, the fraction of businesses that have

[1]Further information can be found in Section 11 of Company Account Guidance

gone into administration is much smaller than that of active companies [5].

Having these data and application-specific challenges in mind, this study proposes and validates a framework for a data-driven tool to facilitate investment decisions. More specifically, the contributions of this paper include:

- Motivation and framing the task of whether a PE firm should invest into a business or not.
- Proposal of a data-driven framework to automate the process of asset screening and thus the recommendation of businesses to invest or not. The framework combines data from several sources to ensure decisions are made holistically considering data related to private equity deals (Unquote data), finances (Fame) and management (Companies House).
- An exploratory data analysis (EDA) of the available data to gain a better understanding about the landscape of PE activities.
- Validation of the data-driven framework for UK-registered businesses by benchmarking its performance for different (i) prediction methods in terms of prediction accuracy and their ability to provide explainable investment recommendations, (ii) class imbalance-handling methods, and (iii) feature sets.

The next section provides more background on how PE investment decisions are done and commonly considered decision criteria, followed by existing research on the application of AI in PE. Section 3 introduces the data-driven framework including an introduction to the available data, data-processing, and the AI methods considered. The EDA of the considered data is presented in Section 4, followed by a validation of the framework in Section 5. Finally, Section 6 concludes the study and discusses future work.

2. Literature Review

Surprisingly, there is little existing research investigating the application of AI to support PE investment decisions. A related and more widely studied domain is the investment in stocks [6, 7]. Although we can learn from that domain when it comes to optimizing PE investment decisions (e.g. in terms of investment criteria), the two types of investment are different and require different modeling techniques.

2.1. Characteristics of good PE investment candidates

A large number of investment criteria could be considered when making a PE investment decision. Which criteria should be considered depends amongst others on the drivers, type of an investment, and the focus of the particular PE business carrying out the investment. We will discuss these considerations and various investment criteria studied in the literature.

Arguably, the most important quantitative criteria are size and profitability-related. This includes Earning Before Interest, Taxes, Depreciation, and Amortization (EBITDA), cash flow and turnover of a business. There are also various qualitative criteria that have been considered when making an investment, such as the company's reputation, the competency of its management team [8, 9, 10], market opportunity, and the product itself. Some of the more popular features from the qualitative criteria include the potential for high profit and exit, the risk of investment and the opportunity for the investors to involve in the prospects of the business [11].

A more recent study by Block et al. [12] found the main criteria for successful investment to be revenue growth, value added of the product or service, management team's track record, international scalability, profitability and the business model. The results also indicated a strong preference for investments in company with reputable investors. This is supported by Dixon and Chong's study [13], using the 'investor rank' algorithm to detect successful early investors. They concluded that current investors in the company are a good indication for the future funding of the target firm. Similarly, Bhat and Zaelit [5] find the first three funding rounds provide sufficient information to later stage investors on the probability of exit with a significant degree of confidence.

Of the financial criteria, positive cash flow [12] and EBITDA are often associated with the highest importance. EBITDA plays an important role in company pricing, as the price for the target firm is determined using a multiple of EBITDA, as long as the EBITDA satisfies some arbitrary minimum requirement [14]. Previous studies show that the average enterprise value of a target company is estimated to be around 7 times of the EBITDA (between 2.7 and 27 times) [15]. This multiple of EBITDA varies based on the firm's and fund's type, and the target company. For example, the EBITDA multiple for transactions involving smaller companies tends to be smaller than the larger transactions [16].

Apart from the financial criteria, empirical literature also states that the calibre of the management has significant influence in the investment decision. Additionally, it has been observed investors' perceptions on industry selection is influenced by market conditions and global financial situation based on company stocks [17]. Gottschlich et al. proposed a decision support system design that uses the crowd's recommendations and investors can use that in their investment decisions and further use it to manage a portfolio [18]. One of the earliest studies of the underlying drivers for success is the survey to identify common criteria selected by the PE firms [19]. The survey results were explained by the authors using an analogy of horse racing, where the market, company, and management are represented by the horse race, horse, and jockey respectively. According to the survey, 10 factors were identified as the indicator for success, covering everything from management and founder team to products, odds, and market. One interesting observation from these results was five of the top ten criteria were related to the management's experience or personality. In other words, the investors should first identify if the 'jockey is fit to ride' - where the track record of the management is assessed to prove the team's ability to react to risk, and

familiarity with the target market or industry.

A more recent study conducted a meta analysis using 31 publications related to the drivers for success in the PE industry [20]. Among the 31 studies, 24 factors were commonly identified across the literature. However, only 8 of these factors were statistically significant and positively correlated to investment performance. These 8 factors were the supply chain integration, market scope, firm age, size of the founding team, financial resources, management experience, founder experience, and the existence of patent protection. On the other hand, 5 of the 24 factors were identified as non-significant predictors. Counter-intuitively, research and development (R&D) cost and competition intensity of the targeted company were part of these 5 non-significant factors.

Another criteria mentioned in the empirical literature is the geographical location of the investing and target firms. In [21], the authors find that PE investors have strong preference for local businesses indicated by the strong negative correlation between the probability of investment and the distance between the target firm and PE firm. This result may be due to the importance of professional networks in PE, and the close proximity between the managers of both firms, which simplifies communication. These findings are also supported by [22] and [23].

2.2. Data-driven methods in PE

In practice, the number of potential companies to invest in is often vast. For instance, according to Companies House, UK's registrar of companies, there are over 4 million private limited companies in the UK [24], hence it is not feasible for PE firms to manually screen all of these companies, even after filtering for key metrics, such as EBITDA and location. Automating part of the screening process could significantly improve the efficiency by showing only relevant companies to deal managers and may lead to more promising investment opportunities, which could have otherwise been overlooked by a human analyst. A survey conducted by von dem Knesebeck [25] reveals that improving the ability to find and execute deals is one of the main motivating factors for 59% of the surveyed VC firms for improving their technology stack. This section provides an overview of the current developments in the PE space.

Dixon and Chong [13] examined the probability of successful investment focusing on private clean-tech businesses.[2] This was done using support vector machine (SVM) based on financial and other qualitative features (reputation of the company), and rank them using Bayesian methods. While this may be a novel approach to the problem, the authors did not provide sufficient evidence on the model performance and feedback.

Bhat and Zaelit [5] apply random forests to predict the probability of PE investment to exit, using data on investor network and funding rounds. Companies are classified into bankrupt or non-bankrupt, and private or public after the

[2] A successful company is defined as one having reached or filed for an initial public offering (IPO) or acquired by another investment firm at a minimum of 1.5 times the investment.

investment. The model successfully predicts the probability of exit using only the funding information from the first three raises with an overall accuracy of 75% and 82% AUC. However, the model performance varies across sectors with the energy sector performing the best in the bankruptcy classification model, and bio-technologies and pharmaceutical sectors performing the best in the private or public company classification model. The authors also apply network analysis to the investor data and observed a positive correlation between the degree of centrality and the probability of exit.

The venture capital (VC) – a sub-field of the PE industry focused on supporting small businesses in their startup phase – has been studied more widely than other fields in PE. Although many of the decision criteria for PE and VC differ, many of the challenges regarding computational approaches are similar. For instance, Zhong et al. [21] propose a probabilistic latent factor model to support investment decisions in the context of startups. The model is further extended with modern portfolio theory [26] to provide recommendations on a personalised portfolio strategy. The reason for using a Bayesian probabilistic method is to overcome the data scarcity and sparsity issue present in this application. The authors predicts the potential returns based on features related to target company's geographical location, number and types of acquisitions and funding rounds, features of the company founders and the product and frequency of news on public media. The experimental study comprised several models and concluded that the proposed probabilistic model achieves a good performance in terms of accuracy; however, more importantly, the study concluded that investors may prefer to invest in a company with more competitors as this may be a sign of booming market.

In general, PE firms are starting to recognise the importance of intelligent data-driven decision making tools. This can be observed via the large investment in human capital (the surge in hiring for data scientist or machine learning engineers) and the rise of FinTech within the PE industry. For example, AlphaSense, an AI-based company that provides investment recommendations derived using natural language processing (NLP) and statistical modelling, or Two Six Capital that perform transaction due diligence using big data and machine learning algorithms. Both of these companies claim to be involved in significant transactions but the actual performance of their solutions are not disclosed.

3. Methodology

Figure 1 presents the proposed data-driven framework for identifying investment opportunities in PE. In summary, data from multiple relevant sources is combined, then pre-processed to be ready for the selection of relevant features, and training and validation of predictive models.

The study uses data from three sources: Unquote, Fame databases and Companies House (CH), where the objective is to make investment recommendations for the UK market. The data from all the sources are merged into a single dataset corresponding to different features for further analysis, which contains data points on companies' financial position, characteristics of the management

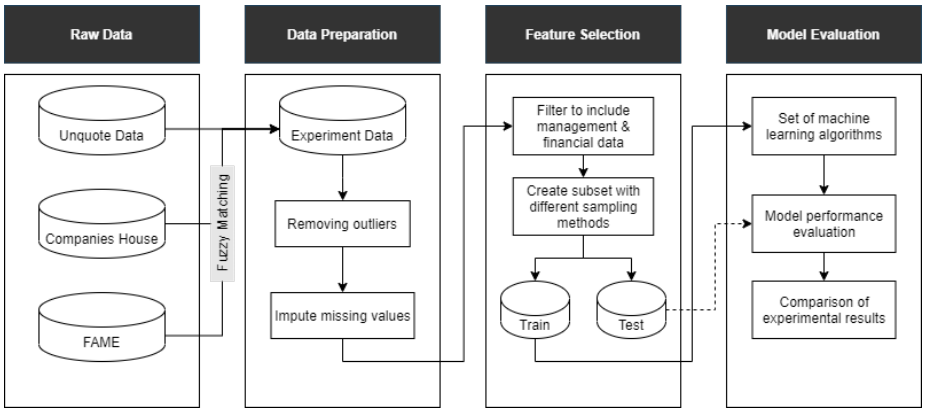


Figure 1: Flow diagram of the proposed data-driven methodology for identifying investment opportunities in PE.

team and investment status. Firstly, the data on PE deals (from Unquote) is merged with the rest of the observations using a fuzzy matching algorithm (discussed in Section 3.1). Secondly, new management related features are engineered from the CH Officer Appointment data, guided by the previous research on the role of management in PE. Finally, six state-of-the-art AI models – Logistic Regression (LR) [27, 28], Decision Tree (DT) [29, 30, 31], Random Forest (RF) [32, 33, 34], k-Nearest Neighbors (kNN) [30, 35, 36], Support Vector Machine (SVM) [37, 38] and eXtreme Gradient Boosting (XG Boost) [39, 40] – are trained and validated using k-fold cross-validation (k=10).

The data provided by Unquote has 18 features and 3248 data points corresponding to each feature for the last 10 years. Each row of the data represents a deal executed with the equity value between £5m to £100m. Some of the important features are described as shown in Table 1. An Exploratory Data Analysis (EDA) was performed using this data to identify the characteristic of the deals, summarising factors like deal size, industry, and distribution of deal over the past 10 years.

3.1. Fuzzy matching

Having decided on set of relevant data sources, the next step is to link them. This can be challenging. In this study, it was not straightforward to link the Unquote data to the CH data as the deal data does not come with unique company identifiers (registration numbers) or standardised company names. For example, 'Cera Ltd' may sometimes be recorded differently as 'Cera', 'Cera - Limited', 'Cera Limited' or 'Cera Holdings'. Due to lack of unique identifiers, the data was ambiguous in nature. Therefore, fuzzy matching, a technique designed to link string records when two strings are not identical, was implemented.

The Levenshtein distance, which computes the similarity of strings based on how many edits are required to make them identical, is the foundation of many string-based comparison algorithms. More precisely, the Levenshtein distance is

Table 1: Description of selected features from the Unquote dataset.

Feature	Data Type	Example
Deal Name	String	Distology
Deal Date	Date	2017-12-03
Country	String	United Kingdom
Deal Value (£m)	String	n/d(50-100m)
	Numerical	16.8
Industry	String	Financial
Business Description	String	Operator of restaurants
Equity Lead	String	NorthEdge Capital
Region	String	Eastern
City	String	London

the minimum number of operations that is needed to make two strings identical, with the operations being insertion, deletion or substitution of a character [41]. More formally, the distance between two strings a and b of length $|a|$ and $|b|$, respectively, is computed as

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise,} \end{cases} \quad (1)$$

where $i \in \{1, 2, ..., |a|\}$ and $j \in \{1, 2, ..., |b|\}$ are indices of the strings a and b, respectively, $\mathrm{lev}_{a,b}(i,j)$ is the distance between the first i and j characters in a and b, and $1_{(a_i \neq b_j)}$ equals to 0 when the two strings are identical and 0 otherwise. The minimum function contains deletion, insertion and substitution elements. As the classic Levenshtein distance does not take into account the length of the strings, a normalized version Levenshtein ratio (LR) [42] is used instead.

The results of the matching are reported in Table 2. Out of 2613 companies in the Unquote dataset, 2081 were matched adhering to the logic described, and 532 were discarded for various reasons including the incorrect labelling of company name, change in company name after transaction, and complication with the parent-subsidiary company structure.

3.2. Data cleaning and engineering

The management related features are derived from the Officer Appointment Snapshot with Resignations data provided by CH. Company officer is a generic term for company directors and secretaries [43]. Thus, when referring to the company officer features, the term 'director features' and 'officer features' are used interchangeably for readability purposes. The following management features are derived:

Table 2: Fuzzy matching results. The number of companies in each fuzzy matching step. The full Unquote dataset information on each PE deal. A number of companies have been involved in more than one PE deal, thus the number of unique target companies is smaller than the number of records in the full dataset. Companies matched with their registration number with high enough confidence represents the number of unique companies that were matched with CH company name with at least a Levenshtein ratio of 90% or with at least 70% Levenshtein ratio and the same city as reported by the Unquote dataset. The Levenshtein ratio is computed after removing company legal status identifiers ('Limited', 'Ltd', 'Group', 'Holdings' etc) and converting all strings to lower case. Percentage represents the fraction of companies in the fuzzy matching step relative to the number of unique companies in the dataset.

Filtering	Number of observations	Percentage, %
Unquote dataset	3248	-
Unique companies in the Unquote dataset	2613	100%
$\geq 90\%$ match confidence	1529	59%
$\geq 70\%$ and $< 90\%$ match confidence	585	22%
$\geq 70\%$ and $< 90\%$ match confidence, matched	552	21%
Companies matched	2081	80%

- Number of active directors
- Number of roles
- Average experience at appointment
- Average number of previous companies
- Average tenure prior to current role
- Cumulative experience at appointment
- Average age at appointment

A business can be incorporated by the founder or company formation agencies, which under the UK law are permitted to guide the process and submit incorporation documents, provide registered office address, help setting up a business bank account and provide ongoing company secretarial support [44]. Thus formation agencies become officers in thousands of companies. For instance, two of the largest officers in the CH Appointments Snapshot dataset, Temple Secretaries Limited and Company Directors Limited, are both linked to over 200,000 records each. This causes a large skew for the average number of previous companies and cumulative experience at appointment features, overshadowing the data on other director experiences. As the presence of formation agency in the company is not expected to provide any important information on the company for PE investors, all institutional officers are removed from the dataset.

A number of the records are discarded from the dataset for the following reasons. Many of the companies received multiple investments in the same year. Because the financial data was joined by deal date, this leads to duplicate records, which would inflate the prediction results, so all duplicate company records are dropped. There are also holding companies, where only one of the

child companies is active and the rest are dormant. As the dataset obtained from the Fame database contains both the parent and child businesses, and the financial data for those companies are very similar, one of the records is dropped to avoid information spill between the test and training sets in the modelling phase.

Lastly, there is a risk of mismatches between the company names provided by the CH and the Unquote database. In the cases where companies are matched incorrectly, the wrong financial and management data is joined with the company, which may result in large outliers. Thus, particular care is taken to ensure all erroneous matches are excluded from the final dataset. The top 2.5% and bottom 2.5% of the matched PE and financial data, calculated by the features that contain the most prominent outliers, i.e. EBITDA, turnover and shareholder funds, are removed from the data set. Outliers in this case suggest that these records are matched with the wrong company or with a large parent company, so removing these records completely is preferable over opting for an outlier substitution method.

The final dataset contains 98385 records and 21 features, of which 814 companies have received PE investment and 97571 companies have not received PE investment. These firms will also be referred to as 'deal companies' and 'no-deal companies'. The final financial features include turnover, turnover growth, EBITDA, EBITDA margin, shareholder funds, number of employees, liquidity, return on shareholder equity (ROSE), profit margin, asset turnover, long term liabilities and minimum EBITDA and EBITDA margin from the past three years. The director features include number of active directors, number of roles in the company, average tenure at appointment, average director age at appointment, average experience and experience in the target company. The features also include the company age in years and a dummy feature, indicating whether a company has received a PE deal previously or not. The summary statistics for all the features are reported in Section 4.

3.3. Modelling: Theoretical background

Section 5 will compare the ability of six AI models – LR, DT, RF, kNN, SVM, XG Boost – to predict the probability of a company receiving PE investment using financial and managerial features. The six models can be divided into two main categories using the complexity-explainability trade-offs. Three models, LR, DT, and kNN, emphasize on the ability to explain the reasoning behind the predicted probabilities to non-technical individuals but at a cost of predictive power potentially. The other three models, SVM, RF and XG Boost, have a higher predictive power but are less explainable.

4. Exploratory Data Analysis (EDA)

This section provides an EDA on key features of the (merged) data source to be used in Section 5 for the experimental study.

Table 3: Descriptive statistics for deal values.

Statistic	Min	1st Qu.	Median	Mean	3rd Qu.	Max	NAs
Value	5.000	7.765	14.000	23.462	30.000	100.000	1506

Table 4: Distribution of transaction with non-disclosed equity value.

Deal Value	Number of Transaction	Percentage
n/d (<25£m)	947	62.9%
n/d (25 - 50£m)	344	22.8%
n/d (50 - 100£m)	215	14.3%

4.1. Deal value

Table 3 summarises the statistics of deal values. It can be seen that deals with lower equity values were significantly more likely to be executed, with the highest frequency between £10 - £15m. On top of that, a constant decrease in the number of deals can be observed as the deal value increases. Out of the 3200 deals, 1506 of them were recorded without disclosing the actual equity values, accounting for over 46% of the data. The values of these deals were estimated using a conservative range by the industry experts. The distribution for these deals is summarised as shown in Table 4. A similar trend was observed in these deals even in the absence of equity values, where deal values and frequency of deals are inversely correlated.

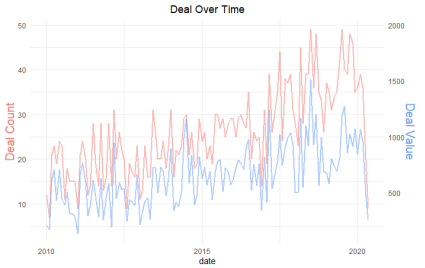
4.2. Growth and cyclicality

Insights on the growth and cyclicality of the PE industry are summarised as shown in Figure 2. A clear upwards trend can be observed on the number of deals over time with some minor fluctuations in Figure 2(a). On the other hand, constant spikes were observed during every March and July in the past 10 years as shown in Figure 2(b), indicating higher activity levels for entry or exit.

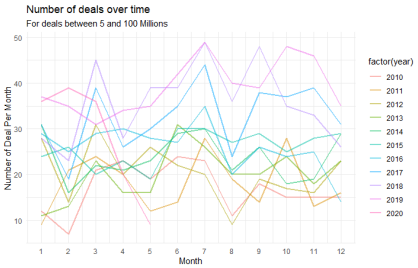
4.3. Company demographics

The demographics of the companies that have received funding in the past including region and industry were explored as follows:

Region. The distribution of companies receiving funding is summarised by location in Figure 3. Figure 3(a) shows that the distribution of deal values across regions was fairly consistent apart from some regions with limited observations, indicating that the PE firms do not necessarily pay a premium for the deal based on the location of the targeted company. However, 40% of the deals in the past 10 years were executed on companies based in London followed by 15% in the South East area of England. This may be due to the availability of resources as the business hub of the country. To note is that this is summary data across

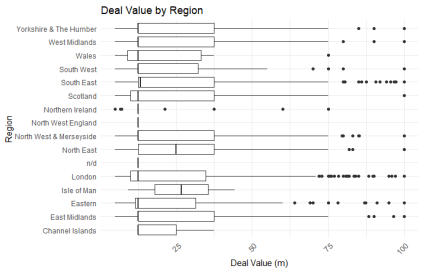


(a) PE deals between 2010 and 2020.

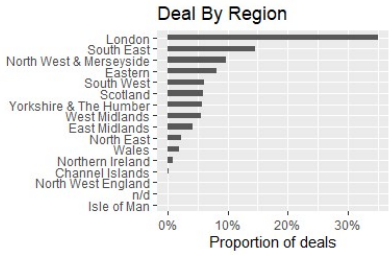


(b) PE deals by month.

Figure 2: PE growth and cyclicality.



(a) PE deals value by region.



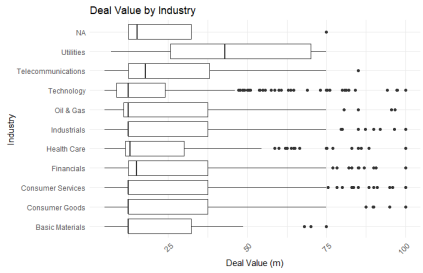
(b) PE deals count by region.

Figure 3: PE deals by region.

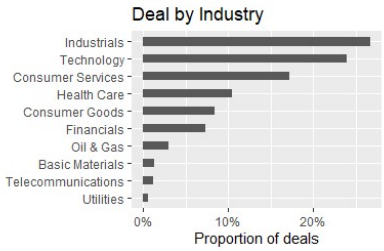
various PE companies; of course, on an individual basis, a PE firm may well limit itself to focus investments in certain regions and/or industries only.

Industry. The distribution of companies by industry is summarised in Figure 4. Figure 4(a), shows that the funding received by companies from the technology and healthcare industry were slightly lower than the companies from other industry. However, despite the lower distribution in deal values, technology and healthcare industry were both observed with the highest number of outliers. The opposite can be observed for companies from the utility industry, as they demonstrated a distribution with higher deal values with no outliers. Next, in terms of the deal count, companies in the industrial sector will most likely receive funding, seconded by companies from the technology sector, which respectively makes up 27% and 24% of the deals between 2010 and 2020.

Region and industry. Some industries might be more prevalent in a specific region due to the availability of resources. Figure 5 shows the distribution of companies that received funding by industry and region across the UK. An interesting observation unique to deals in London is that the technology industry has the highest number of deals by a significant margin, seconded by the consumer services industry. On the other hand, for the remaining deals from other regions, the industrial sector consistently outnumbers the others, with less



(a) PE deals value by industry.



(b) PE deals count by industry.

Figure 4: PE deals by industry.

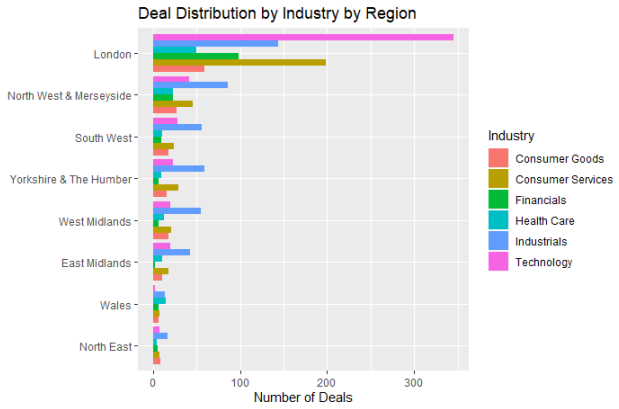


Figure 5: PE deals by industry and regions.

significant differences between the industries.

4.4. Financials

As stated previously, to reflect the financial and management features correctly, they are backdated to the time of investment as opposed to the current state. To ensure our findings are robust, we include confident matches (1796 companies) according to fuzzy matching in the EDA for financial and director data. Table 5 provides some statistics related to different financial features, which will be discussed next.

The average turnover for a company that has received funding is £40m. However, some companies reported extreme values. Examples of such outliers are The Currency Cloud Limited and The Sage Group PLC, who both recorded a turnover of over £2bn at the time of investment. On top of that, the average was also affected by the extreme from the opposite end, as one of the companies recorded a negative turnover of -£2.7m. Due to the lack of legal obligations on small companies as stated in the previous section, about 50% of the matched companies had not filed their turnover at the time of the deal. 54% of the

Table 5: Descriptive statistics for turnover, EBITDA, profit margin and shareholder fund.

Parameter	Minimum	Median	Maximum	Average	% of Missing Values
Turnover (thousand)	-2,721	11,847	2,571,126	40,530	50.01%
EBITDA (thousand)	-77,856	789	111,195	1,813	45.93%
Profit margin	-75.75	4.25	57.11	2.68	57.52%
Shareholder fund (thousand)	-187,719	1,528	1,577,433	8,166	16.81%

Table 6: Descriptive statistics for director data.

Management criteria	Minimum	Median	Maximum	Average
Directors	1.00	4.00	134.00	4.47
Unique positions	1.00	3.00	11.00	3.36
Corporate directors	0.00	0.00	5.00	0.15
Directors with FTSE experience	0.00	0.00	11.00	0.02
Directors with multiple positions	0.00	3.00	22.00	2.91
Average length of tenancy	0.00	3.57	28.00	5.00
Combined years of experience	0.00	15.00	607.00	20.43

companies recorded their EBITDA at the time of the deal, with an average EBITDA of £1.8m. However, more than 50% of the companies reported a negative EBITDA. With regards to profit margin, the median and average values are only slightly in the positive with a reasonable number of companies having a negative profit margin. The majority of companies reported the shareholder fund at the time of deals, with less than 17% in missing values. 57% of companies recorded a shareholder fund of over £1m.

The importance of these financial metrics will later be tested in the model building stage to determine if they should be included in the model.

4.5. Management

Although it is not mandatory for companies to report financial data, they are required by law to provide updated information for the appointment of directors and officers. For this reason, the analysis of management criteria was conducted with no missing values and the results are summarised in Table 6.

Companies that have previously received funding have on average a smaller management team with fewer than five directors covering four different business functions. Moreover, companies tend to be filled with newer directors with an average tenancy duration of five years and 20 years in combined experience. Finally, despite the importance of having a director with experience from larger firms [19], only 14 companies hired directors with previous experience in FTSE100 companies.

5. Experimental Study

Computational asset screening methods conventionally rely on financial and basic company information, such as company age, number of employees and in-

dustry. However, PE investors place as much of an importance on the company's management as on the company's financial performance. This poses challenges for the applications of traditional techniques on the Private Equity (PE) sector due to the difficulty of obtaining and quantifying information about the quality of company directors and management teams. We address these challenges by attempting to quantify a set of proxies for management team quality from officer appointment data, provided by the Companies House. We first conduct a comparative analysis of the performance of six AI models and three methods for dealing with class imbalance (Section 5.1). Hyperparameters of the models were set as suggested in the original papers.

We then perform a more in depth analysis and contrast the predictive power of the company officer and company financial information features (Section 5.2). Generally AI models provide the prediction of the outcome variable. But we have also looked at the contribution of individual features/components in making that specific decision and tried to understand if it also correlate with general way of investing. We have used SHAP values [45] to look at specific feature contribution in making a prediction in PE investment decision making process.

5.1. Comparison of feature sets, algorithms and sampling techniques

Table 7 provides comparison of hold-out sample performance metrics for three different feature sets: financial and basic company information, company officer data, and the combination of both. Compared are the six AI models and three different techniques for dealing with class imbalance – undersampling (at random from the majority class), oversampling with replacement (from the minority class) and SMOTE [46] – to find interesting investment targets for PE investors in a pool of companies. The table reports five different performance metrics: accuracy, precision, F1, ROC and recall rate. The training sample was balanced by either undersampling, oversampling with replacement or SMOTE, but the hold-out sample was left imbalanced with less than 1% of the companies having received investment, to reflect the true state of the market. As it is important for the results in this study to both identify a large set of good investment opportunities, but also have relatively small number of false positives, the further analysis in the followings sections focuses on F1 score, which is the harmonic mean of recall rate and precision, and accounts for the class imbalance in the hold-out set.

Table 7 show the performance of the six AI models, three imbalance-handling (sampling) methods, and three input feature sets using five metrics. Tables 7 and 8 show the average performance of each AI model by sampling techniques and the average performance of all models by input feature sets. It is obvious from the results that there is no clear winning setting. On average, models relying solely on financial features outperform those trained on director features only, across all sampling techniques. However, models that use both financial and director features consistently outperform those trained on one of the feature sets alone (Table 9). The average F1 score across all models for the financial data set, balanced by SMOTE, is 0.09, three times higher than that for director

Table 7: Hold-out sample performance metrics by model, input features and sampling technique.

Sampling technique	Model	Financial Features					Director Features					All Features				
		Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
Undersampling	LR	0.72	0.019	0.036	0.69	0.66	0.62	0.014	0.027	0.64	0.66	0.71	0.019	0.038	0.71	0.71
	RF	**0.75**	**0.025**	**0.049**	**0.78**	**0.80**	**0.72**	**0.020**	**0.039**	0.71	0.69	0.82	0.033	0.063	0.79	0.77
	XGBoost	0.74	0.023	0.044	0.74	0.75	0.71	**0.020**	**0.039**	**0.72**	0.72	0.83	**0.038**	**0.072**	**0.83**	**0.84**
	SVM	0.74	0.021	0.041	0.72	0.70	0.62	0.016	0.032	0.70	**0.78**	**0.86**	0.036	0.069	0.75	0.63
	KNN	0.69	0.020	0.038	0.73	0.77	0.62	0.015	0.030	0.68	0.73	0.70	0.021	0.040	0.74	0.79
	DT	0.70	0.018	0.035	0.69	0.67	0.66	0.016	0.032	0.68	0.70	0.75	0.023	0.045	0.74	0.73
Oversampling	LR	0.67	0.018	0.035	**0.70**	0.73	0.62	0.014	0.027	0.64	0.66	0.71	0.02	0.038	**0.71**	0.71
	RF	**0.99**	**0.455**	0.059	0.52	0.03	**0.99**	0.000	0.000	0.50	0.00	**0.99**	**0.36**	0.058	0.52	0.03
	XGBoost	0.97	0.084	**0.128**	0.62	0.27	0.93	**0.031**	**0.056**	0.60	0.26	0.98	0.18	**0.241**	0.67	0.35
	SVM	0.80	0.024	0.045	0.69	**0.58**	0.73	0.021	0.042	**0.73**	**0.72**	0.94	0.07	0.130	0.76	**0.58**
	KNN	0.97	0.046	0.070	0.56	0.14	0.97	0.021	0.031	0.52	0.06	0.98	0.05	0.068	0.55	0.11
	DT	0.99	0.105	0.096	0.54	0.09	0.98	0.009	0.011	0.50	0.01	0.99	0.12	0.119	0.56	0.12
SMOTE	LR	0.68	0.017	0.034	0.69	**0.70**	0.63	0.014	0.027	0.64	0.65	0.70	0.02	0.037	**0.70**	**0.71**
	RF	0.99	0.161	0.120	0.55	0.09	0.98	0.005	0.006	0.50	0.01	0.99	0.27	0.198	0.58	0.16
	XGBoost	**0.99**	**0.317**	**0.174**	0.56	0.12	**0.99**	**0.028**	0.027	0.51	0.03	**0.99**	**0.49**	**0.287**	0.60	0.20
	SVM	0.86	0.029	0.054	**0.68**	0.50	0.77	0.023	**0.044**	**0.71**	**0.65**	0.96	0.10	0.167	0.74	0.53
	KNN	0.92	0.029	0.052	0.60	0.27	0.92	0.023	0.041	0.57	0.20	0.94	0.04	0.074	0.62	0.30
	DT	0.98	0.061	0.082	0.56	0.13	0.97	0.009	0.012	0.50	0.02	0.97	0.08	0.112	0.60	0.22

Table 8: Average hold-out sample performance metrics by the model and sampling technique.

	Undersampling					Oversampling					SMOTE				
	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
LR	0.682	0.017	0.034	0.680	0.677	0.667	0.017	0.033	0.684	**0.700**	0.668	0.017	0.032	0.677	**0.686**
RF	**0.763**	0.026	0.050	0.759	0.755	**0.991**	**0.271**	0.039	0.510	0.021	0.987	0.144	0.108	0.540	0.086
XGB	0.760	**0.027**	**0.052**	**0.765**	**0.770**	0.961	0.099	**0.142**	0.631	0.295	**0.989**	**0.279**	**0.163**	0.556	0.116
SVM	0.739	0.025	0.047	0.722	0.705	0.825	0.039	0.072	**0.726**	0.627	0.863	0.050	0.088	**0.712**	0.559
KNN	0.670	0.019	0.036	0.718	0.766	0.972	0.039	0.056	0.541	0.103	0.928	0.031	0.056	0.596	0.257
DT	0.703	0.019	0.037	0.701	0.698	0.985	0.078	0.075	0.533	0.074	0.975	0.048	0.069	0.551	0.120

features (0.03). Using both feature sets yields an F1 score of 0.15, almost two times higher than that of financial features alone, and three times higher than the F1 score of director features alone. These results provide evidence that company's financial information alone is a significantly better predictor for the probability of investment than the information on company's officers, but both feature sets strongly compliment each other, and using both financial and management data on average yields almost two times better results.

There is also strong evidence for the benefits of oversampling methods over undersampling methods: models trained on SMOTE all-features dataset on average achieve three times higher F1 score (0.15) than that of under-sampled dataset (0.05), while oversampling with bagging achieves two times higher F1 score (0.11) relative to under-sampling.

Table 8 compares the predictive performance by model and sampling technique across all feature sets. Overall, the best results are achieved by tree-based algorithms in combination with SMOTE, followed by distance-based algorithms and, lastly, linear algorithms. XGBoost on average outperforms all other algorithms across all sampling techniques, achieving the highest average F1 score of 0.163. The results vary vastly across sampling techniques: for under-sampled dataset XGBoost provides a marginal improvement in the F1 score (0.052) of 0.02 points, compared to the next-best model, RF (F1 of 0.05). But for the SMOTE dataset, XGBoost provides an improvement in the average F1 score (0.16) by over 50% compared to the next best model (RF, F1 OF 0.108). XGBoost is followed by RF with the average F1 score of 0.108 and SVM with the average F1 score of 0.088 (SMOTE sampling).

The more granular model comparison by feature set and sampling technique

Table 9: Average hold-out sample performance metrics by input features and sampling technique across all models

	Undersampling					Oversampling					SMOTE				
	Accuracy	**Precision**	**F1**	**ROC**	**Recall**	**Accuracy**	**Precision**	**F1**	**ROC**	**Recall**	**Accuracy**	**Precision**	**F1**	**ROC**	**Recall**
Director	0.66	0.02	0.03	0.69	0.71	0.85	0.02	0.03	0.60	0.34	0.88	0.02	0.03	0.57	0.26
Financial	0.72	0.02	0.04	0.72	0.73	0.91	0.11	0.06	0.59	0.27	0.90	0.10	0.09	0.60	0.30
All	0.78	0.03	0.05	0.76	0.75	0.93	0.13	0.11	0.63	0.32	0.93	0.17	0.15	0.64	0.35

Table 10: Logistic regression coefficients obtained by LR for all input features and SMOTE.

Feature	**Coefficient**	**Standard Error**	**t-statistic**	**p-value**
Constant	4.720	0.09	53.99	0.00
Company age, log	-1.773	0.02	-111.68	0.00
Turnover, log	0.452	0.01	58.85	0.00
Turnover growth	0.002	0.00	13.12	0.00
EBITDA	0.023	0.01	3.87	0.00
EBITDA_margin	-0.029	0.00	-23.57	0.00
Shareholder funds	0.000	0.00	-32.68	0.00
Employees	0.001	0.00	18.74	0.00
Liquidity, log	0.046	0.01	4.62	0.00
ROSE	0.000	0.00	-0.36	0.72
Profit margin	0.015	0.00	17.10	0.00
Asset turnover, log	-0.024	0.00	-25.13	0.00
Long term liabilities	0.001	0.00	5.76	0.00
Min EBITDA	0.279	0.01	22.13	0.00
Min EBITDA margin	0.000	0.00	0.41	0.69
Number of active directors	-0.222	0.00	-59.14	0.00
Number of director roles, log	1.585	0.02	101.20	0.00
Average tenure	0.251	0.00	77.37	0.00
Average age at appointment, log	-0.100	0.00	-58.47	0.00
Number of previous companies, log	0.140	0.01	20.22	0.00
Experience in the company, log	0.069	0.01	5.92	0.00
Average experience at appointment, log	0.394	0.01	35.92	0.00
Number of directors with FTSE experience = 1	0.881	0.03	26.87	0.00
Number of directors with FTSE experience = 2	-0.488	0.04	-11.34	0.00
Number of directors with FTSE experience >= 3	0.416	0.08	5.28	0.00

in Table 7 provides additional insights about algorithms' ability to capture complex patterns and benefit from additional information. XGBoost achieves a high F1 score of 0.29 (all features, SMOTE balancing), but the performance of this algorithm is closely related to the amount and complexity of the data available. The capabilities of XGBoost are best seen for large datasets, i.e. over-sampled and using all features. XGBoost is an advanced algorithm, capable of capturing complex nonlinear relationships, but requires large amounts of data to be effective, and heavily overfits the smaller undersampled dataset. With increasing complexity, XGBoost performance improves significantly. The findings are similar to other related studies. For instance, [47] compare XGBoost, RF, SVM and LR on a credit scoring problem and find similar patterns in performance: XGBoost outperforms RF, SVM and LR in that order, measured by accuracy. But the authors found the difference between the best and worst performing algorithm less extreme: XGBoost provided an improved upon the accuracy between LR by only 6%. We find there is very little change in the predictive power in LR, no matter the feature set or sampling technique.

Table 10 reports the LR model coefficients and test statistics. Almost all

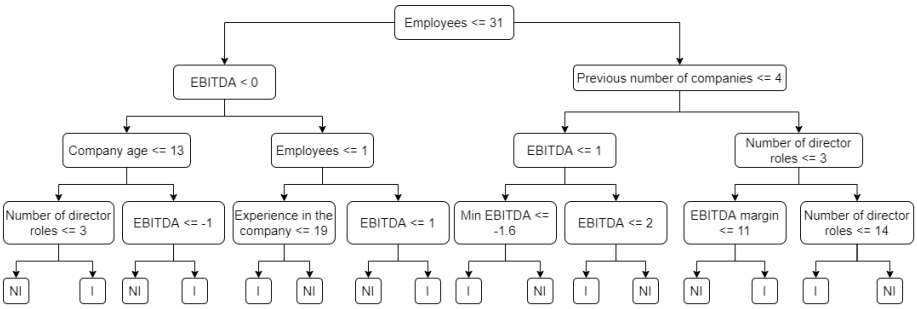


Figure 6: Decision tree obtained by DT for all input features and SMOTE.

features are statistically significant with the exception of ROSE and minimum EBITDA margin. The variables with the highest economic significance include company age, EBITDA, turnover, min EBITDA and number of director roles. LR achieves an F1 score between 0.027 and 0.038 but the algorithm fails to capture complex patterns in the data set, and providing additional information does not improve the performance. Whereas for other models the predictive power increases with increasing complexity of sampling technique and larger data set. An exception, however, is the director feature data set. For models using only company officer information, the predictive power is relatively low across all sampling techniques with F1 score varying between 0 and 0.056.

DT achieves a slightly higher performance in terms of F1 (0.112), reporting similar features as the most important ones; this can be observed from the derived decision tree as shown in Figure 6 using the combined input feature set. Among the first decision levels, there are EBITDA, number of previous companies, company age, number of employees and number of director roles. The decision tree also reveals several patterns that align with intuition. For example, the model suggests to invest in (i) small, young companies with high number of directors and negative EBITDA, or (ii) companies with large number of employees with directors that have experience in a number of other companies and small EBITDA. If the directors do not have the aforementioned experience, higher number of director roles (i.e. breadth of experience over depth of experience) are considered of similar importance.

Relative to XGBoost, non-boosted tree algorithms, i.e. RF and DT, the performance is more consistent: the models have less of a tendency to overfit for small datasets, but cannot capture as many complexities in the larger datasets as XGBoost. RF and SVM produce relatively consistent performance across different feature sets and sampling techniques, with relatively good results for both small and large data sets, and performance improvements when more information is provided. The F1 score of SVM varies between 0.032 and 0.167. RF performs well for small data sets, being the top algorithm for both undersampled financial and director feature sets. An exception is the over sampled director feature data sets, where RF achieves a low F1 score of 0. Overall, our findings are consistent with previous literature. For instance, [48] evaluate 170

classifiers from 17 algorithm families across 121 data sets; Random Forest (RF) was found to be the most likely best is classifier, achieving 90% accuracy in 84% of the datasets, followed by SVM, neural networks and boosting ensembles. In [49], SVM, neural networks and DT classifiers are applied to a credit scoring problem and it has been found that SVM achieves identical classification accuracy as the other models but with fewer input features. It should, however, be noted that the runtime of SVM increases exponentially with the volume of input data, so in practice SVM models are not feasible for similar problems. We observed significantly longer training times for SVM for over-sampled data sets, relative to the other algorithms.

5.2. Explainability

Understanding the model decision process and the role of individual features is imperative for setting a good baseline benchmark for future PE asset screening models and advancing the theoretical knowledge base of PE decision making process. Examining certain feature interactions can be useful for bringing important details to the attention of investment analysts, when conducting the subsequent business analysis following the asset screening stage. This section is dedicated to the interpretation of the modelling results and the discussion on the importance of individual features and their interactions. In order to gain meaningful insights about the model decision process, it is crucial to have a good base model. As it was observed in Section 5.1, the XGBoost model trained on the SMOTE data set achieves the best predictive power out of all the algorithm combinations we studied. Therefore, the following section focuses on the analysis of this particular model only. We use Shapley Additive Explanations (SHAP) values to estimate feature importance and SHAP interaction values to analyse the feature dependencies.

Figure 7 shows the top twenty features ordered by their importance in predicting the probability of receiving PE investment in the hold-out sample, where importance is measured by the SHAP values. The top five features are company age, EBITDA, the number of employees, the number of directors in the company and long term liabilities. The bottom features, that are not included in the plot, are number of directors with previous experience in a FTSE 500 company and the min EBITDA of the past three years. This reiterates the previous statement about PE placing more importance on the management than the company's financial performance: while profitability margins are among the features with the lowest importance, management-related features, such as number of directors, average experience prior to the acquired company, the management team's average tenure and number of director roles have relatively high SHAP values. The results described in the previous section demonstrate that financial and basic company information attain better model predictive power relative to management features alone, but according the SHAP values, most management features carry higher importance in predicting the probability of the company receiving PE investment than the financial features. This discrepancy arises from feature interactions: management features alone are not strong predic-

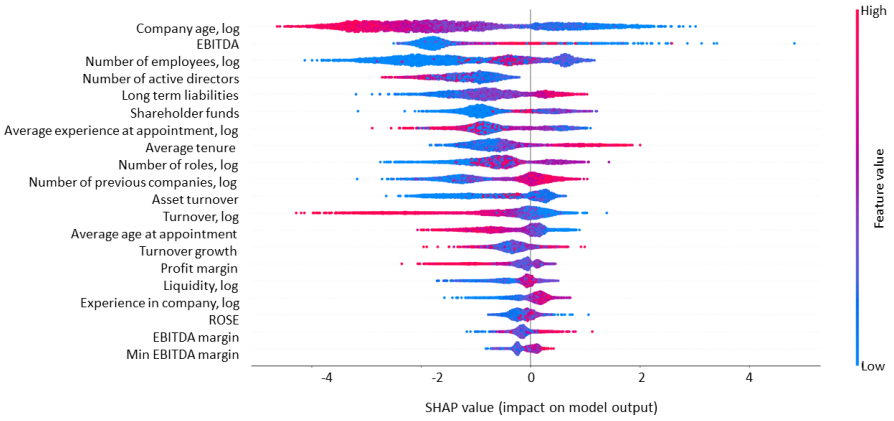


Figure 7: Feature importance with feature effects on model output, SMOTE

tors of probability for investment, but combined with financial information, the director features become essential.

Additionally, Figure 7 shows the relationships between the probabilities of receiving PE investment and the feature values. The figure plots hold-out observations by feature and SHAP value. The horizontal axis represents the SHAP value of impact on model output for each feature shown on the y-axis, and the colour of each observation indicates the relative value of each observation, with red being high and blue low. Negative SHAP values indicate a reduction in the probability of investment in the model decision making process, while positive SHAP values increase it. Several features stand out with strong relations between their relative values and the impact on model output, particularly, company age, EBITDA, long term liabilities, average tenure, number of companies with experience as a director, turnover and the average age at appointment.

Company age (Figure 8), the feature with the highest impact on model output magnitude, has a strong negative correlation with the probability of receiving PE investment. High age strongly impedes the chance of receiving investment, whereas young age increases it. In addition, according to SHAP, the negative impact for age is stronger than the positive, respectively, company being relatively old is a better predictor of no PE investment than young age being a predictor of investment. PE firms tend to invest in medium maturity companies with proven demand and product, so recently established businesses may carry more risk than desired, but older companies may be less appealing for several reasons, such as business model, lack of innovation, financial success and owner's attitude towards external investors. If a business is successful and shows potential signs of positive future returns, it is highly likely a PE firm will have expressed the desire to invest in the company. In this case the business owners

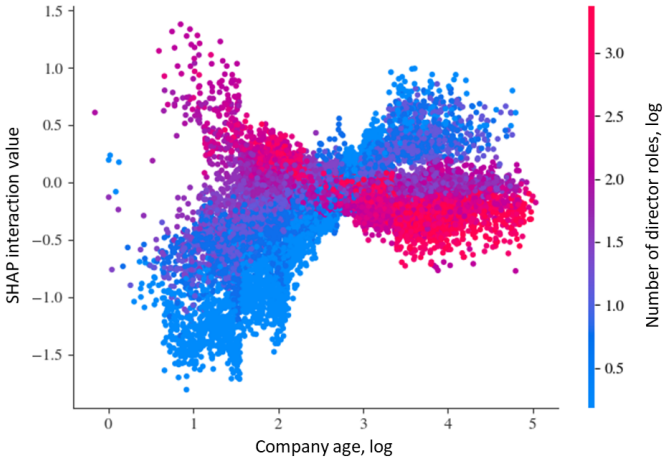


Figure 8: SHAP interaction values for company age vs number of director roles

will either accept the offer (and the business would have been classified as an investment company) or the owners decline the offer due to the preference for full ownership or other reasons. Thus, the older the company gets, the smaller the probability for PE investors to find 'hidden gems', and the company's age becomes a predictor for investment by itself.

The financial features that proxy company size, i.e. EBITDA and turnover, appear to be more important than the relative profitability features, such as EBITDA margin or profit margin. This reveals the inherent flaws of AI models that rely solely on financial information: other than company size features, there is little information to be extracted from the profitability indicators, as PE investors do not consider these as their primary point of interest. It must be noted, however, that this may not hold for leveraged buyout (LBO) firms and distressed company investments, where poor financial performance is the primary condition for investment and less importance is placed on the management team, as it is likely to be replaced after the investment. The EBITDA feature shows an interesting relationship between the feature variable and SHAP values, that related to this observation. Observations with high EBITDA are concentrated around zero SHAP value, while low values are more present at the extreme ends. The PE firms concentrating on distressed businesses are likely to invest in low EBITDA companies, whereas firms looking for less risky investments are looking for higher EBITDAs.

Management features can be divided into three groups: experience (the average years of prior experience as a director, the average number of previous companies, average age at appointment, years of experience in the company and number of directors with experience in a FTSE 500 company), skill diversity (the number of active directors, number of different director roles) and

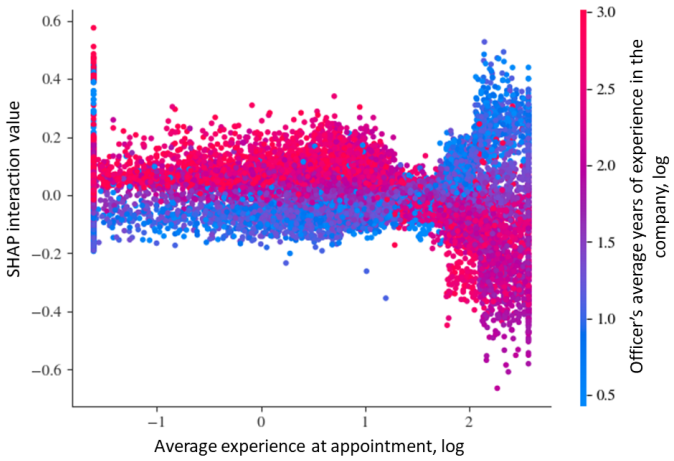


Figure 9: SHAP interaction values: average experience at the time of appointment and the average number of years in the company

turnover (tenure of the officer team members). According to Figure 7, the most important management features include features from each of the group, with the top three being the number of active directors, the average years of previous director experience as company officer, and the average director tenure in the team. There is a positive bias towards repeated founders as a company's success is closely related to the number of companies an owner has founded previously. The average number of companies the directors have held a management role at can be associated with more start-up experience and proven ability to start/scale up companies. The number of employees and number of directors are not correlated (0.08), meaning that higher number of directors in the company is not purely an implication of a larger company. High SHAP values are associated with low number or director roles and low number of employees or vice versa.

The average tenure and the number of companies the directors had experience prior to being appointed have similar strong positive correlation with the probability of PE investment. Average officer tenure in the company may be a proxy for the quality of the company's internal culture and founder's dedication to the company. High officer turnover may signal issues within the business. The number of roles within a company ranks the ninth, showing that very high or low number of director roles in a company decrease the probability for investment, but the middle ranges increase it. This illustrates the significance of breadth of experience over multiple officers with similar experience.

For young to medium aged companies (approximately 3-8 years) it is preferable to have at least 4-5 directors, but for older companies above the age of 10 years, having no more than roughly 5 directors increases the probability for investment (Figure 8). This can be explained with the breadth versus depth ar-

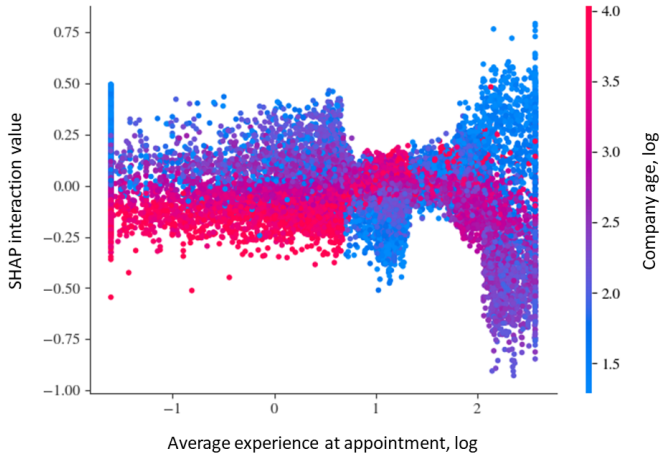


Figure 10: SHAP interaction values: average experience at appointment and the company age

gument. For younger companies the inherent risk can be diversified with wider variety of skills, whereas for older companies when the risk is less substantial, it is more important to have a small team with extensive experience in the business and industry. A similar pattern can also be observed in Figure 10, which shows the SHAP interaction values between the company age and director experience. The probability of receiving PE investment for a young company below the age of 5 years is significantly higher when the directors on average have had at least 5 years of experience prior to the appointment. For less experienced teams this probability is higher if the company is at least 7 years old. However, as discussed previously, companies aged above 20 years have negative SHAP values associated with it, meaning companies above this age are less appealing investment targets. This can also be seen in Figure 9, which the SHAP interaction values between the logged average years of experience at the time of appointment and the logged average experience in the company. According to the SHAP values, the probability of PE investment increases for companies with directors of at least 5 years of experience as officers in different companies prior to joining the target business and low average number of years in the target company. That is, PE investors have a strong preference for young companies, founded by experienced founders. If the officers are less experienced, i.e. have between 0 and 5 years of experience as officers, the probability of investment increases if the officers have spent roughly 9 or more years in the business. First time founders may be riskier bet for PE investors, thus they need to spend more time building the business and proving its success compared to experienced investors for PE.

6. Conclusion and Future Work

This was the first study that investigated the application of AI to facilitate holistic PE investment decision-making. After motivating the problem of investing into a business or not as a classification problem, we proposed a data-driven framework comprising the (fuzzy-matched) merger of 3 data sources (containing information about financial and managerial properties of businesses and previous PE deals), a data preparation pipeline, a feature selection and class imbalance method, and finally a model evaluation and explanation stage. Following an EDA of the available data to gain a better understanding and assess the application of AI, we conducted a comparative analysis on the model performance using three different feature sets across six different AI algorithms and three class imbalance-handling techniques. We then used SHAP values to examine the importance of financial and management features in the decision-making process and model predictive power. In summary, we found that:

- Fundamental company and financial information has more predictive power of PE investment than managerial data related to a company alone, but the best results are achieved using both feature sets.
- Shap values showed that company age has the highest impact on PE investment. High age strongly impedes the chance of receiving investment, whereas young age increases it. The followed second important trait is EBITDA.
- The XGBoost model trained on the SMOTE data set achieves the best predictive power out of all the algorithm combinations studied.
- The best results are achieved by tree-based algorithms in combination with SMOTE, followed by distance-based algorithms and, lastly, linear algorithms.

This study can be extended in various ways. We validated the proposed framework for the United Kingdom but it would also be interesting to understand if and how drivers for investment decisions vary in other countries. In addition to management and financial data about companies, one can, of course, consider various other information to consider in investment decisions, such as ESG (Environmental, Social, and Governance) performance of a company. However, there is a lack of (open-source) data sources that contain such information (e.g. ESG) for all registered companies but it would be interesting to understand how this information relates to company valuation and PE investment decisions. Finally, while we considered binary decisions about investing/not investing in a company, it would be interesting to investigate whether a more fine-grained assessment (e.g. Likert scale for investment opportunities) leads to a more robust model; however, obtaining such labels for training is likely to be a time-consuming manual task.

Appendix A. Results

Table A.1: Model evaluation metrics by model and input feature set, trained on under-sampled dataset. Train and test set panels represent the average 5-fold cross validation score. Training set is balanced by under-sampling the majority (investment) class. Hold-out panel represents the metric obtained by a model retrained on the training and test data combined.

		Train					**Test**					**Hold-out**				
		Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
Director features	LR	0.66	0.65	0.67	0.66	0.70	0.62	0.015	0.029	0.66	0.69	0.62	0.014	0.027	0.64	0.66
	RF	1.00	1.00	1.00	1.00	1.00	**0.71**	**0.021**	**0.042**	**0.75**	0.78	**0.72**	**0.020**	0.039	**0.71**	0.69
	XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.69	0.019	0.037	0.72	0.75	0.71	0.020	**0.039**	0.72	0.72
	SVM	0.74	0.70	0.76	0.74	0.83	0.61	0.017	0.032	0.71	**0.81**	0.62	0.016	0.032	0.70	**0.78**
	KNN	0.79	0.74	0.80	0.79	0.87	0.59	0.015	0.030	0.68	0.78	0.62	0.015	0.030	0.68	0.73
	DT	1.00	1.00	1.00	1.00	1.00	0.66	0.015	0.030	0.65	0.64	0.66	0.016	0.032	0.68	0.70
Financial features	LR	0.69	0.69	0.70	0.69	0.71	0.68	0.018	0.034	0.69	0.69	0.72	0.019	0.036	0.69	0.66
	RF	1.00	1.00	1.00	1.00	1.00	**0.76**	**0.025**	**0.049**	**0.77**	**0.78**	**0.75**	**0.025**	**0.049**	**0.78**	**0.80**
	XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.74	0.024	0.046	0.75	0.76	0.74	0.023	0.044	0.74	0.75
	SVM	0.78	0.75	0.79	0.78	0.84	0.69	0.020	0.039	0.73	0.77	0.74	0.021	0.041	0.72	0.70
	KNN	0.81	0.79	0.82	0.81	0.85	0.69	0.019	0.037	0.71	0.74	0.69	0.020	0.038	0.73	0.77
	DT	1.00	1.00	1.00	1.00	1.00	0.70	0.017	0.034	0.68	0.66	0.70	0.018	0.035	0.69	0.67
All features	LR	0.73	0.72	0.73	0.73	0.74	0.70	0.019	0.037	0.71	0.72	0.71	0.019	0.038	0.71	0.71
	RF	1.00	1.00	1.00	1.00	1.00	0.81	0.035	0.068	**0.83**	**0.84**	0.82	0.033	0.063	0.79	0.77
	XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.82**	**0.036**	**0.069**	0.83	0.84	0.83	**0.038**	**0.072**	**0.83**	**0.84**
	SVM	0.83	0.80	0.83	0.83	0.87	0.79	0.029	0.055	0.78	0.77	**0.86**	0.036	0.069	0.75	0.63
	KNN	0.83	0.79	0.84	0.83	0.89	0.69	0.021	0.040	0.75	0.82	0.70	0.021	0.040	0.74	0.79
	DT	1.00	1.00	1.00	1.00	1.00	0.74	0.023	0.044	0.74	0.74	0.75	0.023	0.045	0.74	0.73

Table A.2: Model evaluation metrics by model and input feature set, SMOTE. Train and test set panels represent the average 5-fold cross validation score. Training set is balanced by over-sampling the minority class by SMOTE. Hold-out panel represents the metric obtained by a model retrained on the training and test data combined.

		Train					**Test**					**Hold-out**				
		Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
Director features	LR	0.68	0.67	0.70	0.68	0.74	0.63	0.01	0.03	0.65	0.67	0.63	0.01	0.03	0.64	0.65
	RF	1.00	1.00	1.00	1.00	1.00	**0.98**	0.01	0.01	0.50	0.01	0.98	0.01	0.01	0.50	0.01
	XGB	**1.00**	**1.00**	**1.00**	**1.00**	0.99	**0.98**	0.00	0.00	0.50	0.00	**0.99**	**0.03**	0.03	0.51	0.03
	SVM	0.82	0.77	0.83	0.82	0.91	0.73	**0.02**	**0.04**	**0.74**	**0.75**	0.77	0.02	**0.04**	**0.71**	**0.65**
	KNN	0.97	0.94	0.97	0.97	**1.00**	0.92	0.02	0.04	0.56	0.20	0.92	0.02	0.04	0.57	0.20
	DT	1.00	1.00	1.00	1.00	1.00	0.97	0.01	0.01	0.50	0.02	0.97	0.01	0.01	0.50	0.02
Financial features	LR	0.71	0.70	0.72	0.71	0.75	0.67	0.02	0.03	**0.70**	**0.73**	0.68	0.02	0.03	**0.69**	**0.70**
	RF	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	0.19	0.14	0.55	0.11	**0.99**	0.16	0.12	0.55	0.09
	XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.33**	**0.21**	0.58	0.16	**0.99**	**0.32**	**0.17**	0.56	0.12
	SVM	0.86	0.82	0.87	0.86	0.92	0.83	0.02	0.05	0.66	0.50	0.86	0.03	0.05	0.68	0.50
	KNN	0.98	0.96	0.98	0.98	1.00	0.92	0.03	0.05	0.60	0.26	0.92	0.03	0.05	0.60	0.27
	DT	1.00	1.00	1.00	1.00	1.00	0.98	0.06	0.09	0.56	0.15	0.98	0.06	0.08	0.56	0.13
All features	LR	0.74	0.73	0.75	0.74	0.78	0.71	0.02	0.04	0.72	**0.74**	0.70	0.02	0.04	**0.70**	**0.71**
	RF	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.21	0.14	0.55	0.11	0.99	0.27	0.20	0.58	0.16
	XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.53**	**0.28**	0.59	0.19	**0.99**	**0.49**	**0.29**	0.60	0.20
	SVM	0.92	0.90	0.93	0.92	0.96	0.93	0.07	0.12	**0.77**	0.61	0.96	0.10	0.17	0.74	0.53
	KNN	0.98	0.95	0.98	0.98	1.00	0.93	0.05	0.08	0.66	0.38	0.94	0.04	0.07	0.62	0.30
	DT	1.00	1.00	1.00	1.00	1.00	0.97	0.06	0.09	0.58	0.18	0.97	0.08	0.11	0.60	0.22

Table A.3: The average model evaluation metrics by input feature set for under - sampled dataset.

	Train					**Test**					**Hold-out**				
	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
Director	0.86	0.85	0.87	0.86	0.90	0.65	0.02	0.03	0.69	0.74	0.66	0.02	0.03	0.69	0.71
Financial	0.88	0.87	0.88	0.88	0.90	0.71	0.02	0.04	0.72	0.73	0.72	0.02	0.04	0.72	0.73
All	**0.90**	**0.89**	**0.90**	**0.90**	**0.92**	**0.76**	**0.03**	**0.05**	**0.77**	**0.79**	**0.78**	**0.03**	**0.05**	**0.76**	**0.75**

Table A.4: The average model evaluation metrics by input feature set for over - sampled (SMOTE) dataset.

	Train					**Test**					**Hold-out**				
	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
Director	0.91	0.90	0.92	0.91	0.94	0.87	0.01	0.02	0.57	0.28	0.88	0.02	0.03	0.57	0.26
Financial	0.92	0.91	0.93	0.92	0.94	0.90	0.11	0.10	0.61	0.32	0.90	0.10	0.09	0.60	0.30
All	**0.94**	**0.93**	**0.94**	**0.94**	**0.96**	**0.92**	**0.16**	**0.13**	**0.65**	**0.37**	**0.93**	**0.17**	**0.15**	**0.64**	**0.35**

Table A.5: Average evaluation metrics by model for under - sampled dataset.

	Train					**Test**					**Hold - out**				
	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
LR	0.69	0.69	0.70	0.69	0.72	0.67	0.02	0.03	0.69	0.70	0.68	0.02	0.03	0.68	0.68
RF	1.00	1.00	1.00	1.00	1.00	**0.76**	**0.03**	**0.05**	**0.78**	**0.80**	**0.76**	0.03	0.05	0.76	0.76
XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.75	0.03	0.05	0.77	0.78	0.76	**0.03**	**0.05**	**0.77**	**0.77**
SVM	0.78	0.75	0.79	0.78	0.85	0.70	0.02	0.04	0.74	0.78	0.74	0.02	0.05	0.72	0.70
KNN	0.81	0.77	0.82	0.81	0.87	0.65	0.02	0.04	0.72	0.78	0.67	0.02	0.04	0.72	0.77
DT	1.00	1.00	1.00	1.00	1.00	0.70	0.02	0.04	0.69	0.68	0.70	0.02	0.04	0.70	0.70

Table A.6: Average evaluation metrics by model for SMOTE dataset

	Train					**Test**					**Hold - out**				
	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall	Accuracy	Precision	F1	ROC	Recall
LR	0.71	0.70	0.72	0.71	0.75	0.67	0.02	0.03	0.69	**0.71**	0.67	0.02	0.03	0.68	**0.69**
RF	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.14	0.10	0.54	0.08	0.99	0.14	0.11	0.54	0.09
XGB	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.29**	**0.17**	0.56	0.12	**0.99**	**0.28**	**0.16**	0.56	0.12
SVM	0.87	0.83	0.87	0.87	0.93	0.83	0.04	0.07	**0.72**	0.62	0.86	0.05	0.09	**0.71**	0.56
KNN	0.97	0.95	0.97	0.97	**1.00**	0.93	0.03	0.06	0.61	0.28	0.93	0.03	0.06	0.60	0.26
DT	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.97	0.04	0.06	0.55	0.11	0.97	0.05	0.07	0.55	0.12

References

[1] G. W. Fenn, N. Liang, S. Prowse, The private equity market: An overview, Financial Markets, Institutions & Instruments 6 (4) (1997) 1–106.

[2] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016.

[3] W. Wang, W. Chen, K. Zhu, H. Wang, Emphasizing the entrepreneur or the idea? The impact of text content emphasis on investment decisions in crowdfunding, Decision Support Systems 136 (2020) 113341.

[4] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[5] H. S. Bhat, D. Zaelit, Predicting private company exits using qualitative data, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2011, pp. 399–410.

[6] F. G. Ferreira, A. H. Gandomi, R. T. Cardoso, Artificial intelligence applied to stock market trading: A review, IEEE Access 9 (2021) 30898–30917.

[7] Q. Gao, D.-L. Xu, An empirical study on the application of the evidential reasoning rule to decision making in financial investment, Knowledge-Based Systems 164 (2019) 226–234.

[8] D. A. Shepherd, A. Zacharakis, Conjoint analysis: A new methodological approach for researching the decision policies of venture capitalists, Venture Capital: An International Journal of Entrepreneurial Finance 1 (3) (1999) 197–217.

[9] B. Clarysse, M. Knockaert, A. Lockett, How do early stage high technology investors select their investments?, Tech. rep., Vlerick Business School (2005).

[10] M. Broere, Decision-making in private equity firms: An empirical study of determinants and rules, Springer Science & Business Media, 2013.

[11] L. Feeney, G. H. Haines Jr, A. L. Riding, Private investors' investment criteria: Insights from qualitative data, Venture Capital: An International Journal of Entrepreneurial Finance 1 (2) (1999) 121–145.

[12] J. Block, C. Fisch, S. Vismara, R. Andres, Private equity investment criteria: An experimental conjoint analysis of venture capital, business angels, and family offices, Journal of corporate finance 58 (2019) 329–352.

[13] M. Dixon, J. Chong, A Bayesian approach to ranking private companies based on predictive indicators, AI Communications 27 (2) (2014) 173–188.

[14] E. Stafford, Replicating private equity with value investing, homemade leverage, and hold-to-maturity accounting, Homemade Leverage, and Hold-to-Maturity Accounting (May 20, 2017).

[15] A. Ljungqvist, M. P. Richardson, The investment behavior of private equity fund managers, Tech. rep., New York University (2003).

[16] B. Puche, R. Braun, Deal pricing and returns in private equity, The Journal of Alternative Investments 21 (3) (2018) 70–85.

[17] H. Dincer, U. Hacioglu, E. Tatoglu, D. Delen, A fuzzy-hybrid analytic model to assess investors' perceptions for industry selection, Decision Support Systems 86 (2016) 24–34.

[18] J. Gottschlich, O. Hinz, A decision support system for stock investment recommendations using collective wisdom, Decision Support Systems 59 (2014) 52–62.

[19] I. C. MacMillan, R. Siegel, P. Subba Narasimha, Criteria used by venture capitalists to evaluate new venture proposals, Journal of Business Venturing 1 (1) (1985) 119–128.

[20] M. Song, K. Podoynitsyna, H. Van Der Bij, J. I. Halman, Success factors in new ventures: A meta-analysis, The Journal of Product Innovation Management 25 (1) (2008) 7–27.

[21] H. Zhong, C. Liu, J. Zhong, H. Xiong, Which startup to invest in: A personalized portfolio strategy, Annals of Operations Research 263 (1-2) (2018) 339–360.

[22] R. Florida, D. F. Smith Jr, Venture capital formation, investment, and regional industrialization, Annals of the Association of American Geographers 83 (3) (1993) 434–451.

[23] W. W. Powell, K. W. Koput, J. I. Bowie, L. Smith-Doerr, The spatial clustering of science and capital: Accounting for biotech firm-venture capital relationships, Regional Studies 36 (3) (2002) 291–305.

[24] Companies House, Companies register activities statistical release (2020). URL https://www.gov.uk/government/publications/companies-register-activities-statistical-release-2019-to-2020/companies-register-activities-2019-to-2020

[25] P. von dem Knesebeck, What's Your VC's Tech Stack? Results From A Survey Of Early Stage Venture Capital Funds (2017). URL https://www.bfp.vc/whats-your-vcs-tech-stack-results-from-a-survey-of-early-stage-venture-capital-funds/

[26] E. J. Elton, M. J. Gruber, S. J. Brown, W. N. Goetzmann, Modern portfolio theory and investment analysis, John Wiley & Sons, 2009.

[27] M. Maalouf, M. Siddiqi, Weighted logistic regression for large-scale imbalanced and rare events data, Knowledge-Based Systems 59 (2014) 142–148.

[28] A. Blanco-Justicia, J. Domingo-Ferrer, S. Martinez, D. Sanchez, Machine learning explainability via microaggregation and shallow decision trees, Knowledge-Based Systems 194 (2020) 105532.

[29] J. R. Quinlan, Decision trees and decision-making, IEEE Transactions on Systems, Man, and Cybernetics 20 (2) (1990) 339–346.

[30] L. Jiang, Z. Cai, D. Wang, S. Jiang, Survey of improving k-nearest-neighbor for classification, in: Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007), Vol. 1, IEEE, 2007, pp. 679–683.

[31] V. Marriboyina, R. Lokanatha, A comparative study on decision tree classification algorithms in data mining, International Journal of Computer Application in Engineering, Technology and Sciences 2 (2) (2010) 24–29.

[32] L. Breiman, Bagging predictors, Machine learning 24 (2) (1996) 123–140, available at: https://doi.org/10.1007/BF00058655. doi:10.1007/BF00058655.

[33] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[34] S. Roy, D. Rathee, A. Chowdhury, K. McCreadie, G. Prasad, Assessing impact of channel selection on decoding of motor and cognitive imagery from MEG data, Journal of Neural Engineering 17 (5) (2020) 056037.

[35] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, Z. Xu, Missing value estimation for mixed-attribute data sets, IEEE Transactions on Knowledge and Data Engineering 23 (1) (2010) 110–121.

[36] X. Zhu, L. Zhang, Z. Huang, A sparse embedding and least variance encoding approach to hashing, IEEE Transactions on Image Processing 23 (9) (2014) 3737–3750.

[37] W. S. Noble, What is a support vector machine?, Nature biotechnology 24 (12) (2006) 1565–1567.

[38] A. Fan, M. Palaniswami, Selecting bankruptcy predictors using a support vector machine approach, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, Vol. 6, 2000, pp. 354–359.

[39] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, p. 785–794.

[40] F. Nwanganga, M. Chapple, Improving performance, Practical Machine Learning in R (2020) 341–366.

[41] V. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet physics doklady 10 (8) (1966) 707–710.

[42] A. Marzal, E. Vidal, Computation of normalized edit distance and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (9) (1993) 926–932.

[43] UK Government, Life of a company part 2: Event driven filings (2019). URL https://www.gov.uk/government/publications/life-of-a-company-event-driven-filings/life-of-a-company-part-2-event-driven-filings

[44] Companies House, Company accounts guidance (2019). URL https://www.gov.uk/government/publications/life-of-a-company-event-driven-filings/life-of-a-company-part-2-event-driven-filings

[45] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768–4777.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[47] L. Munkhdalai, T. Munkhdalai, O.-E. Namsrai, J. Lee, K. Ryu, An empirical comparison of machine-learning methods on bank client credit assessments, Sustainability 3 (2019) 699.

[48] M. Fernández-Delgado, E. Cernadas, S. Barro, D. Amorim, Do we need hundreds of classifiers to solve real world classification problems?, The Journal of Machine Learning Research 15 (1) (2014) 3133–3181.

[49] C.-L. Huang, M.-C. Chen, C.-J. Wang, Credit scoring with a data mining approach based on support vector machines, Expert Systems with Applications 33 (4) (2007) 847–856.