Finance 705

Research Proposal: Data Science in Private Equity

Connor Robert McDowall

June 14, 2021

1 Abstract

This research proposal aims to validate and extend the contributions made by Block et al (2019). They identify seven investment criteria crucial for investment decision making in private equity. We explore if data science can improve screening efficiency in investment due diligence for private equity fund managers when assessing investment and exit opportunities. We propose forming numeral and categorical proxies for the investment criteria from private company data stored in PitchBook. Subsequently, we propose the implementation of the data science process proposed by Aurélien Géron (2017). We will train three supervised learning algorithms to make predictions on investment/exit opportunities. These models are Multi-nominal Logistic Regressions, Random Forests, Multi Layer Perceptrons (MLP). The proposed contributions aim to validate the investment criteria, validate the use of PitchBook for research purposes, and show evidence data science can inform investment due diligence and create efficiencies in screening for investments.

Contents

List of Figures

List of Tables

2 Introduction

Private Equity (PE) is an alternative asset class with similar characteristics to hedge funds. Private equity funds are investment vehicles usually identified by four characteristics. Firstly, they are privately organized, pooling capital from several parties. Secondly, professional investment managers administer the fund. Their incentives are performance-based including compensation and significant carry in the fund. Thirdly, they are inaccessible to the public. Lastly, they operate externally to securities regulation and registration requirements. A private equity fund is managed by general partners (GP) who manage limited partners' (LP) investments in the fund. LPs make passive investments with little to no control in the fund's operations. Private equity funds typically charge a 2% annual fee and 20% performance fee on the fund's annual return. They raise capital through private offerings and pursue investment strategies in private markets based on the funds mandate to generate returns for investors (Brav, Jiang, Partnoy, and Thomas, 2008). Data science combines scientific methods, maths, statistics, specialized programming, advanced analytics, AI, and even storytelling to uncover and explain the business insights buried in data (IBM, 2021b). The number of data science applications are increasing in most industries but there is slow uptake in private equity. This creates opportunities to use emerging technologies to add value in operational and investment processes, exemplified through case studies featuring BCG (2019), Blackstone (2020) and the NZ Super Fund. Section 3 evaluates the prior literature related to private equity and decision processes by fund managers. Section 4 frames hypotheses to extend the contributions by Block et al (2019). In particular, this section explores if data science can improve screening efficiency in investment due diligence for private equity fund managers when assessing investment opportunities? Section 5 outlines the variables of interests with investment criteria and investment decision/exit opportunities, the dependent and independent variables respectively. Furthermore, explanations outline the derivation of investment criteria from the database PitchBook, limitations with the database, and contingency plans on the provision of poor quality data from PitchBook. Section 6 conveys how this research proposal will follow conventional data

science processes proposed by Aurélien Géron (2017), and implement the three forms of supervised learning algorithms: Multi-nominal Logistic Regression, Random Forests, and Multi-Layer Perceptrons. Section 7 concludes with the key contributions and a few suggestions for future research.

3 Literature Review

Prior literature emphasizes generated returns, comparisons to public markets, and value created from private equity. The presence of cyclicality in PE returns differs according to fund type and is consistent with the conjecture that capital market segmentation contributes to private equity returns (Cavagnaro, Sensoy, Wang, and Weisbach, 2019). Institutional investors returns are not from chance alone, but skill leads to outperformance when selecting private equity investors (Ang, Chen, Goetzmann, and Phalippou, 2018). The adaptation of stochastic discount factor valuation methods to evaluate performance for venture capital generalized the Popular Market Equivalent (PME) method to reflect risk-free rates and public returns found abnormal performance (Korteweg and Nagel, 2016). Evidence of differences in skill and exit styles among venture partners investing at the same VC firm at the same time estimates human capital is two to five times more important than a VC firm's organizational capital in explaining performance (Ewens and Rhodes-Kropf, 2015). Classification of risks and post-investment actions inform agency and hold-up problems are important to contract design and monitoring (S. N. Kaplan and Strömberg, 2004). The analysis of firm and VC characteristics, in combination with value-increasing investments post-IPO for both VC's and underlying companies, is an efficient solution to information problems (Iliev and Lowry, 2020). Harris et al (2014) found buyout performance consistently exceeds the public markets (S&P 500) by 3% annually, calculated using the Burgiss data set. The performance in Cambridge Associates and Preqin datasets is qualitatively consistent with Burgiss but lower in Venture Economics. The determinant of leverage in buyouts is variation in economy-wide credit conditions. Higher deal leverage is associated with higher transaction prices

and lower buyout fund returns. This suggests that acquirers overpay when access to credit is easier (Axelson, Jenkinson, Strömberg, and Weisbach, 2013). Investments in innovation, measured by patenting activity, informs one form of long-run activity. Based on 472 LBO transactions, there is no evidence that LBOs sacrifice long-term investments (Lerner, Sorensen, and Strömberg, 2011). Phalippou (2020) finds evidence private equity performance does not exceed public markets after considering carry and other factors. The above literature eludes to the presence of multiple factors and investment criteria when making investment decisions. Block et al (2019) explore investment criteria with an experimental conjoint analysis of private equity fund-types to inform how investments are made. There has been a comprehensive investigation on the effects of PE financing in corporate and entrepreneurial finance. Empirical analysis in precedent literature finds evidence of improvements to operating performance from PE investment (S. N. Kaplan and Stromberg, 2009) and public market outperformance ((Ang et al., 2018),(Braun, Jenkinson, and Stoff, 2017), (Harris, Jenkinson, and Kaplan, 2014), (S. N. Kaplan and Schoar, 2005), (S. N. Kaplan and Sensoy, 2015), (Phalippou and Gottschalg, 2009), (Robinson and Sensoy, 2013)). The exploration of PE investments across fund types and size yield consistent results ((S. Kaplan, 1989), (Chemmanur, Krishnan, and Nandy, 2011)). Selection and treatment effects ascribe to increased performance ((Bengtsson and Sensoy, 2011), (Bernstein, Giroud, and Townsend, 2016), (Rin, Hellmann, and Puri, 2013), (Puri and Zarutskie, 2012)). The active investment nature of PE enables portfolio companies the provision of value-added activities, either direct or indirect. Direct benefits include access to coaching or networks. Indirect benefits include certification effects to third parties ((Bottazzi, Da Rin, and Hellmann, 2008), (Gompers and Lerner, 2001), (Hellmann and Puri, 2002), (Korteweg and Sorensen, 2017), (Josh Lerner, 1995)). Portfolio company selection, and the capacity to add value through financial, governance and operational engineering, are the skillsets emphasized by PE investors. Contrary to the importance of investment selection, there is very little literature exploring investment selection and decision making by private equity managers. PE managers expend considerable resources in evaluating and screening investment opportunities ((S. N. Kaplan and Stromberg, 2001),

(Gompers, Kaplan, and Mukharlyamov, 2016)). Their investment screening and selection process reviews many companies while only investing in a select few. Gompers et al (Gompers et al., 2016) reports for every hundred investment opportunities, the average PE investor conducts thorough due diligence on 15, enters agreements with eight, and eventually closes fewer than four. Empirical challenges associated with isolating the effect of different company characteristics contribute to the lack of empirical evidence surrounding investment criteria. Block et al (2019) are one of the first groups to investigate the investment criteria of PE investors, given decision making in PE is often debated ((Gompers and Lerner, 2001), (S. N. Kaplan and Strömberg, 2004)). The use of observational data is not feasible as observing investor preferences between two identical companies that vary in predetermined characteristics is not possible. Adopting similar methodologies to Bernstein et al (2017), Block et al (2019) compares decision making across different investor types using a large-scale conjoint analysis of 19,474 screening decisions by 749 PE investors through contacting 15,600 investment professionals listed in PitchBook. The conjoint analysis enables a more accurate representation of actual decision making as captures decisions making trade offs between criteria. Block et al (2019) required participants to make a series of assessments on a set of discrete company attributes. In particular, these attributes are: (1) profitability, (2) revenue growth, (3) track record of management team, (4) reputation of current investors, (5) business model, (6) value-added of product/service, and (7) international scalability. Every participant needed to evaluate multiple companies which differ only in the specifications of the above attributes and recommend investment decisions. A multi-level logistic regression model evaluated and compared the importance of different investment criteria, enabling criteria comparisons across investor types. Lerner et al (Josh Lerner, Schoar, and Wongsunwai, 2007) identifies there are likely differences in decision making between investor types with a broader perspective on investing behaviour underdeveloped (Hellmann, Schure, and Vo, 2013). Block et al (2019) investigates analysis with greater granularity to explore decisions by different investor types. In particular, investor types explored in this analysis are, (1) family offices, (2) business angels, (3) venture capital funds, (4) growth

equity funds, and (5) leveraged buyout funds. Firstly, Block et al (2019) identify the relative importance of PE investors' investment criteria. In order of importance, revenue growth, value-added (product/service), and management team track record are the most important criteria. Internationally scalability, current profitability, business model, and reputation of existing investors are relevant but of lower importance. Secondly, Block et al (2019) compare the importance of the respective investment criteria across different investor types. They provide systematic empirical comparison methods of these differences and find family offices, growth equity funds and leveraged buyout funds prefer profitability over revenue growth. Venture capital funds and business angels prefer revenue growth over profitability. These findings imply discrepancies in the risk profiles between investor types. In the case of family offices, the results align with the objective of a family office to preserve wealth in order to maintain financial and social standing. The above literature informs the investment criteria to considered when evaluating investment decisions in PE.

4 Hypotheses Development

4.1 Research Question

This research proposal aims to extend the contributions from Block et al (2019). Gompers et al (Gompers et al., 2016) inform the closure of fewer than four out of hundred investment opportunities. The wrong investment decisions may have serious consequences for both fund returns and manager reputation. Kaplan et al (2001) and Gompers et al (2016) reiterate PE managers expend considerable resources evaluating and screening investment opportunities. Block et al (2019) identified several investment criteria integral to investment decisions: (1) profitability, (2) revenue growth, (3) track record of management team, (4) reputation of current investors, (5) business model, (6) value-added of product/service, and (7) international scalability. Data science can automate manual processes and make predictions considering complex interactions between numerical and categorical variables. Data science methodologies considering the variables proposed

by Block et al (2019), after surveying PE managers and their risk profiles, may reduce screening time, reduce due diligence costs, and identify profitable investments aligning with investment mandates. Subsequently, this research proposal will explore:

Can data science improve screening efficiency in investment due diligence for private equity fund managers when assessing investment opportunities?

In layman's terms, improving screening efficiency relates to accurately predicting suitable companies to invest. This process supports company selection when using traditional screening processes and considers the investment criteria proposed by Block et al (2019). Depending on the success of screening models, these practices may replace existing screening methods, adding value to both investment due diligence processes and PE fund managers.

4.2 Hypotheses

The proposition of the below hypotheses aims to investigate the research question:

$$H_0 : \text{Data science models do not predict suitable investment targets.} \quad (1)$$

$$H_1 : \text{Data science models do predict suitable investment targets.} \quad (2)$$

Suitable investment targets are companies that either currently, or are predicted to, align with Block et al (2019). Data science methodologies will also test if the investment criteria proposed by Block et al (2019) after surveying PE managers identify suitable investment targets and/or if there are other unexplained contributing factors/interactions.

5 Data

5.1 Variables: Inputs

Block et al (2019) implemented a two-step process to evaluate the screening criteria. Firstly, prior research informs an investment criteria long-list ((Bernstein, Korteweg, and

Laws, 2017), (Franke, Gruber, Harhoff, and Henkel, 2008), (Puri and Zarutskie, 2012)). Secondly, Block et al (2019) conducted 19 expert interviews with PE investors across Europe and the US, identifying the most relevant criteria After their analysis, they discovered the relative importance of PE investors' investment criteria across multiple investor types. These criteria are, (1) profitability, (2) revenue growth, (3) track record of management team, (4) reputation of current investors, (5) business model, (6) value-added of product/service, and (7) international scalability. One must highlight the definition of business models comes from Amit et al (2001). Block et al (2019) outlined the attributes and attribute levels for the above investment criteria, used in the conjoint analysis, in figure 1. The following subsections include the descriptions of these attributes (figure 1). This proposition aims to explore these criteria paired with common financial, operational and categorical variables displayed in table 1.

5.1.1 Revenue Growth

Revenue growth describes the average yearly revenue growth over the last years. This is a categorical variable with four designations: 10% p.a., 20%p.a., 50%p.a., and 100%p.a. These growth rates will be considered over one, three and five year time periods, assigned to the closest category. Additionally, revenue growth will be included as a numerical variable for comparison purposes.

5.1.2 Value-added (Product/Service)

Value-added services (product/services) describes the value added to the customer from the product or service. Low value represents a marginal improvement (e.g., cost reduction or service quality), whereas high value represents significant improvements. This is a categorical variable with three designations: low, medium, and high. Value-added is a difficult variable to measure. However, using sentiment analysis with Natural Language Processing (IBM, 2021a) with non-financial data (e.g. social media mentions, web traffic, news features and reviews) would enable the categorization of value-added services.

5.1.3 Management Team Track Record

Management team track record describes whether the management team has a relevant track record (e.g. industry and leadership experience). This is a categorical variable with three designations: none of them, some of them, all of them. Multi-nominal logistic regression will consider executive experience and education to create the above categorical variables (Edgar and Manz, 2017).

5.1.4 International Scalability

International scalability describes the difficulty of scaling internationally, in terms of the time and investments needed. This is a categorical variable with three designations: easy, medium and difficult. Multi-nominal logistic regression will consider various features (e.g., industry classification, committed capital, market presence, years since founding etc.) to create the required categorical variables (Edgar and Manz, 2017).

5.1.5 Profitability

Profitability describes the current profitability of the company, a categorical variable with three designations: not profitable, breakeven, and profitable. Multi-nominal logistic regressions will consider the following financial values to form the above categorical variables: Earnings Before Interest, Tax, Depreciation and Amortization (EBITDA); Earnings Before Interest and Tax(EBIT); Net Operating Performance After Tax (NOPAT); Return on Assets (ROA); and Return on Equity (ROE). Additionally, EBIT, EBITDA, NOPAT, ROA, and ROE will be included as numerical variables for comparison.

5.1.6 Business Model

Business model describes the key focus of the company based on prior research (Amit and Zott, 2001) pertaining to four designations: (1) Lock-in, (2) Innovation-centered, (3) Low cost, and (4) Complimentary. The Lock-in model keeps customers attracted and 'locked in', having high switching costs for customers, which prevent them from changing to other providers. The Innovation-centered model offers innovation in the form of new technology,

products or services. The Low cost model focusses on reducing costs for customers for already existing products or services. The Complimentary model bundles multiple goods and services to generate more value for customers. This is also a difficult variable to derive. Business descriptions would be input into the text classification functionalities in Natural Language Processing (IBM, 2021a) to categorize models designations. This will be difficult.

5.1.7 Current Investors

Current investors describes the types of investors, if any. This is a categorical variable with three designations: no other current external investors, other current external investors - unfamiliar, and other current external investor - Tier 1. Tier 1 investors are reputable investors. Investor relationships can be modelled using the mathematics behind graph theory and network analysis from geographical applications ((Curtin, 2018), (Faudree, 2003)). Modelling the strength in investor relationships and investment history, in combination with multi-nominal logistic regression analysis, will categorize the necessary designations. However, this will also be difficult.

5.1.8 Year

The consideration of a yearly designation (e.g., 2016, 2017 etc.) related to the collection of investment criteria proposed by Block et al (2019) will inform time-series analysis.

5.2 Variable: Output(s)

Block et al (2019) explored the importance of investment criteria through multi-level logistic regressions models. The investment decision is binary: 0 if no investment, 1 if investment. Multi-level logistic regressions account for both nested investment decisions and multi-level effects. This proposal will explore a similar outcome of an investment decision (0 if no investment, 1 if investment). Additionally, the exploration of three exit outcomes based on investment criteria will contribute to the validation of an investment target. These outcomes are, (1) IPO, (2) Acquisition, and (3) Bankruptcy/failure. In

lyman's terms, identify the likely exit outcomes and suitable time to exit a company for a PE manager. Ross et al (2021) explore this phenomena with models trained on a different set of features using a different dataset (Crunchbase, 2021). The inclusion of the desired outcomes in the datasets enables the most appropriate algorithms for this proposal (Section 6.5). It must be highlighted their findings are not published in top ranked journal, they don't consider key financial or operation variables in their feature selection, and lacks rigorous comparisons to traditional empirical methods (e.g., logistic regressions) to cross-validate results. It is poor quality research. This research proposal will explore the above exit outcomes using investment criteria considered by PE investors and cross-validate results with logistic regressions where possible.

5.3 Sources

5.3.1 PitchBook

PitchBook is financial data and software company that provides thousands of professional's comprehensive data on private and public market information (PitchBook, 2021). Block et al (2019) identifies PitchBook as one of the most comprehensive databases in entrepreneurial finance, regularly used for PE related research ((S. N. Kaplan and Lerner, 2016),(Paglia and Harjoto, 2014)). Disclosed information from limited partners, filings of national regulators and other available information are the main contributors to the database. PitchBook has advantages over alternative databases as reports information on investment teams and contact details in addition to information on the investment entity (Brown, Harris, Jenkinson, Kaplan, and Robinson, 2015). The database records comprehensive data on companies, investors, deals, M&A, LPs, funds, financials, advisors, professionals, debt & lenders. In particular:

- Companies of various designations (Publicly traded, Pre-IPO, PE-backed, Startups/Stealth etc.)
- Deal information (Bankruptcies, IPOs, PIPEs, LBO, VC Investments etc.)
- Financial information (calculation transparency, balance sheets, cash flow state-

ments, income statements, consensus information, deal multiples, financial ratios, fundamentals etc.)

Data, both time-series and cross-sectional, is accessible using application programming interfaces (API) or direct downloads to excel formats (e.g., xlsx etc.) Industry widely adopt the platform as the product has high levels of granularity for data science applications. PitchBook continues to grow as industry and PitchBook employees continue to contribute to the platform. An itemized illustration, of the size and scope of available data, as at 03/06/2021 follows:

1. **Deals**: 1,540,549 deals with 45 deals types, evaluate deal histories, get key information and deal multiples, access pre and post money valuations, explore series terms and stock information.

2. **Companies**: 3,096,933 (private), 58,362 (public), get key information, explore financing history, evaluate financials and filings, view executives and board members, follow non-financial metrics.

3. **Financials**: Financials and estimates summary, analyse key metrics, explore balance sheets, income statements, cash flows, ratios & multiples.

PitchBook is an appropriate data source as Block et al (2019) surveyed investors from this database, and it provides all the information required to derive the set of variables described in Section 5.1.

5.4 Limitations: PitchBook

The main contributors to PitchBook are disclosed information from limited partners, filings from national regulators and other publicly available information. Additionally, there is self-selection bias as private companies elect to disclose information-related to their companies. This research proposal is unable to provide descriptive statistics on the data available from PitchBook as requires a service subscription. The derivation of Business Model and Value-added categorical variables will be difficult as require an

implementation of a complex algorithm. However, it is feasible. The aforementioned limitations may create poor quality datasets when taking global perspectives. This research proposal will focus on the North American market in an attempt to minimize data issues as data pertaining to this market is the most complete. If the data is unsuitable for analysis after these considerations, we will explore the alternative datasets and sources in Section 5.5.

5.5 Alternative Data Sources

The consideration of alternative data sources will form contingency plans if data issues persist with PitchBook. A consortium of databases is necessary to construct the variables of interest described in MergerMarket (2021) and CB Insights (2021) have a comprehensive M&A database on relevant deals. Crunchbase (2021) contain investment professional, identification and non-financial information on early stage companies. Preqin (2021) includes comprehensive information on PE managers. In addition to the above databases, liaising and partnering with local and global PE managers may help source the required data to create variables. However, PitchBook is more comprehensive than the consortium of alternatives as contains all the required information for investment criteria variable construction. The research proposal may be put on hold until better quality data comes to market on the basis both PitchBook and the consortium of alternatives fail to provide the required data.

6 Methodology

The research methodology will follow the conventional data science process proposed by Aurélien Géron (2017). This process is: (1) Get the "Big picture". (2) Get the Data. (3) Discovery & Visualization. (4) Data Preparation. (5) Model Selection & Training. (6) Model Tuning. (7) Presentation. (8) Launch, Monitor, and Maintain System (Omitted). We explore each step sequentially from Section 6.1 to Section 6.7. and highlight the mathematics pertaining to the methodology is listed in the Appendix (Section 8.1).

6.1 Problem Scoping

6.1.1 Problem Framing

Problem scoping involves three processes: framing the problem, selecting performance measures, and checking data assumptions. Sections 2, 3 and 4 frame the objective of this research proposal. Currently no other solutions or prior literature on using the investment criteria proposed by Block et al (2019) exist to train data science models with the objective of informing and increasing the efficiency of investment screening and selection for PE fund managers.

6.1.2 Performance Measure Selection

Performance measures evaluate the accuracy of machine learning models to validate predictions. The computation of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) will inform predictability. Both are cost functions and measure the distance between vectors. RMSE calculates the standard deviation of errors between the observed values ($y^{(i)}$) and values predicted by the model ($h(\mathbf{x}^{(i)})$. This performance measure is the preferred the algorithms discussed in Section 6.5. The number of instances in the dataset (m) is comparable to a datapoint with combinations of time-series and cross-sectional data in conventional empirical datasets. $\mathbf{x}^{(i)}$) is a vector of all feature values for the ith instance. In layman's terms, the investment criteria (feature values) are described in Section 5 for the ith company (instance).

6.1.3 Checking Data Assumptions

This research proposal builds on the investigation Block et al (2019). The numerical and categorical variables suggested in 5 are appropriate. The outcomes of this research will inform an end-to-end methodology that do not rely on other systems or processes. Addressing the data limitations in Section 5 enable the implementation of this research proposal.

6.2 Data Acquisition

Firstly, there will be an initialization of a research workspace using the technologies itemized in the Appendix (8.2) and best practise proposed by Wilson et al (2016). The workspace enables a Python implementation, distributed by Anaconda, using several core modules: Numpy, Pandas, Matplotlib, Scikit-learn, and Tensorflow. Microsoft Visual Code (IDE), complete with Git functionalities, will facilitate software development. The hardware in use will be a MacBook Pro (2019) with macOS Mojave as the chosen operating system. The use of IBM Watson and AWS technologies depends on hardware capabilities and the computational complexity of this methodology. PitchBook, a subscription-based product, will provide the raw data pertaining to private companies. Subsequently, collaboration with subscription holders is necessary. A screening for US companies sorts US company-related information into one data source. This sheet contains the relevant information to derive the required numerical and categorical investment criteria. Downloading the results from the screening grants access to the data. This is the matrix for feature values for all instances. PitchBook grows iteratively as there are over 3 million private companies registered on the database. It is important to automate a data pipeline to retrain models and provide up to date information. This will be accomplished through controlling PitchBook's excel application programming interface (API) with a custom python module. If the above data limitations related to PitchBook persist, industry collaboration enables the access to relevant data from their internal and external sources. Additionally, the division of the feature matrix ($\mathbf{X}$) into three sets is necessary to mitigate bias in both model selection and data snooping. Random sampling will form three subsets: training, validation and testing respectively.

6.3 Discovery & Visualization

Following on from data acquisition, it is important to get a general understanding of the data prior to manipulation and preparation. Exploratory analysis will take place on the training set investigating a number of features pertaining to the dataset. This analysis will explore geographical visualizations per US State to show concentrations of

private entities, correlations between numerical and categorical variables pertaining to unprocessed investment criteria and explore the correlations between combinations of numerical and categorical variables.

6.4 Data Preparation

A series of transformation functions will be written for reproducibility on updated datasets and test variation in data transformations. This will take the form of a transformation pipeline to apply to the training set feature matrix ($\mathbf{X}$). The pipeline will first incorporate cleansing functionalities to: remove missing variables, isolate the required variables in the feature matrix in order to derive investment criteria in Section 5, Numerical revenue and profitability values are converted to the designated categorical investment criteria proposed by Block et al (2019). Value-added (Product/Service) and Business Model categorical variables will require Natural Language Processing techniques to derive the attribute levels within these investment criteria. The use of sentiment analysis with non-financial data (Section 5) using Tensorflow's Word2Vec & Seq2Seq tutorials will derive Value-added (Product/Services) categories. The use of NLP's text classification functionalities with company descriptions (Section 5) will categorize business models. Variations in generalized multi-level logistic regressions (Section **??**) using the relevant data to the desired investment criteria (Section 5) will categorize the variables needed for International Scalability, Management Team Track Record and Current Investors investment criteria. The categorization of desired outcomes (outputs) for screening processes (investments and exits) will be binary (1 for each of a present outcome (investment, IPO, acquisition, bankruptcy/failure), 0 otherwise). After, missing variables will be filled with median values on a case-by-case basis in order to include critical instances. Thirdly, the transformation pipeline will convert text and categorical attributes to numerical values stored as SCiPy sparse matrices using scikit-learn's OneHotEncoder function. Lastly, the transformation pipeline will implement feature scaling to rescale all input attributes to the same scale using scikit-learn's StandardScaler function. Standardization subtracts the mean value and divides by the variance so the resulting distribution has unit vari-

ance. This procedure does not bound values to a specific range (unlike normalization) but is much less affected by outliers. Rescaling is required to optimize algorithm performance.

6.5 Model Selection & Training

Data preparation is the most difficult section of the methodology. Methods become elementary (my dear Watson) after data preparation. Machine learning methods have a reputation for being 'blackboxes' with decision making opaque to users. This generally creates adoption issues which may extend to PE managers given the lack of understanding in how the algorithms function. Subsequently, this research proposal suggests three supervised learning, model-based methodologies: (1) Multi-nominal Logistic Regression, (2) Random Forests, and (3) Multi Layer Perceptrons (MLP). Supervised machine learning algorithms include desired outcomes (labels) in the training sets and are the most transparent to PE managers e.g., these investment criteria contribute to this investment or exit decision. These models enable predictions across both time-series and cross-sectional contexts. These algorithm require nested mathematical functions. We include their expression in the Appendix (8.1).

6.5.1 Multi-nominal Logistic Regression (Softmax Regression)

?? Logistic regression estimates the probability that an instance belongs to a particular class. This research proposal will consider the investment criteria of each instance to determine the probability of investment and exit outcomes. The use of this algorithm is intuitive as enables cross-validation with the logistic regressions performed by Block et al (2019). This proposal suggests using two multi-nominal logistic regression, generalizations to support one investment decision and three exit outcomes respectively. This algorithm computes a score, then estimates the probability of each class by applying the normalised exponential. After calculating the scores for every class for the instance $\mathbf{x}$, the probability $\hat{p}_k$ that the instance belongs to class k is calculated by running the scores through a softmax function. This function computes the exponential of every score then

normalizes them. The softmax regression classifier predicts the class with the highest estimate probability. The models aims to estimate a high probability for the target class (and low probabilities for the other classes) through the minimization of a cross entropy cost function. The computation of a gradient vector from the cross entropy cost function enables optimisation techniques, in this instance gradient descent, to find the parameter matrix Θ that minimizes the cost function. Therefore the investment or exit decision. Sckit-learn's LogisticRegression function applies this algorithm with the equation pertaining to the method in the Appendix ().

6.5.2 Random Forests

The second algorithm is a Random Forest, an ensemble of decision trees. Decision trees are an algorithm suitable for classification tasks. They make predictions based on a branching structure and are relatively simple. This research proposal suggests decision trees can estimate the probability an instance (company) belongs to a particular investment decision or exit opportunity. Estimation starts at the root node. Each node explores two outcomes e.g., revenue growth is less than 20% p.a. or 20% and greater. The algorithm will divide the training instances based on the binary criteria. A node will have three attributes: (1) samples, (2) value, and (3) gini. Sample counts how many instances the node applies to e.g, 100,000 companies. Value informs the number of instances per class applies to the node e.g, 40,000 invested, 60,000 not invested. Gini is an impurity measure with purity (gini = 0) representing all training instances it applies to belong to the same class. There are two forms of impurity measure: gini and entropy. Decisions at nodes form boundaries, forming partitions of instance groupings. Decision trees can continue to form new nodes based investment criteria until it settles on the max depth or each leaf node is 'pure'. Subsequently, a decision tree can estimate the probability that an instance belongs to an investment decision or exit opportunity. Firstly, it traverses the tree to find the leaf node for this instance, then returns the ratio of training instances of class k in this node. Requesting a prediction will return the class with the highest probability at this node. The Classification And Regression Training (CART) algorithm

trains the decision trees. Firstly, it splits the training set in two subsets using a single feature k and threshold t_k, searching for the pair producing the purest subsets (weighted by their size). This process is recursive until either the user-defined max depth is reached or there are no further splits that will reduce impurity. The 'optimal solution' is difficult to find as the optimal tree is a NP-Complete problem, requiring O(exp(m)) time causing intractability for fairly small trees. Gini impurity measures leads to faster computations but entropy measures tend to produce more balanced tree. Random forests are ensembles of decisions trees, combining combinations of decision trees trained using the same training algorithms with varying subsets of the training data. This process forms a diverse set of classifiers with predictions, when aggregated, This proposal suggests using Scikit-learn's RandomForestClassifier to implement ensembles of decisions trees, applying the above process, to compute the equations in the Appendix (8.1.4). Random forests are the most appropriate for PE-related decision making as they are the most intuitive and transparent of the three proposed supervised machine learning algorithms.

6.5.3 Multi Layer Perceptron (MLP)

Artificial Neural Nets (ANN) are versatile, powerful, and scalable. They sit at the heart of deep learning as frequently outperform other machine learning algorithms on large and complex problems. This research proposal suggests implementing two sets of multi-layer perceptron, a form of ANN, to predict investment decisions and exit opportunities from investment criteria. A linear threshold unit (LTU) feeds the weighted sum of input values ($z = \mathbf{w}^T \cdot \mathbf{x}$) into a step function ($h_w(\mathbf{x}) = step(z)$). A perceptron is a single layer of LTUs where each LTU is connected to every input. Perceptrons are suitable for classification as output the positive investment decision or exit opportunity if a threshold is met. Perceptrons utilize a training algorithm assessing the strength of connections between perceptrons while considering errors. A perceptron is fed one training instance at a time, making predictions for each instance. For every output LTU that produced a wrong prediction, it re-enforces the connection weights using the perception learning rule (Appendix (8.1.5)) from the inputs that would have contributed to the right pre-

diction. A Multi Layer Perceptron is composed of one LTU input layer, multiple LTU hidden layers and an output LTU layer. The step functions in each LTU are replaced by a logistic or ReLU function ($\sigma(z) = \frac{1}{1+exp(-z)}$ or $ReLU(z) = max(0, z)$ respectively) to enable gradient descent for optimisation. A shared softmax function replaces the individual activation functions in the output layer to enable exclusive classification. In this instance, the classification of investment decisions, or exit opportunties, from investment criteria. Tensorflow's DNNClassifier function facilitates the implementation of MLPs in this proposal.

6.5.4 Evaluation

This proposal will utilise performance measures outlined in Section 6.1.2 and cross-validation techniques with validation sets to inform the prediction proposed models. Transparency decreases and complexity increases with Random Forests, Multi-nominal Logistic Regression and Multi Layer Perceptrons (MLP) respectively.

6.6 Model Tuning

Each model has a set of hyper parameters integral to performance. This proposal will conduct grid search, randomized search, error analysis and evaluations using test sets to find the best combination of hyper parameters to optimise model performance.

6.7 Solution Presentation

This research proposal will inform the feasibility of data science applications in private equity to assist with investment screening due diligence. The above methodology will be converted to a custom python package distributed on the open-source Python Package Index (2021) containing source code, package documentation, this research proposal, accompanying dissertation and technical guides to inform proposed algorithms. Lastly, demonstrations will be made to private equity managers.

7 Conclusions

7.1 Contributions

This proposal is a very difficult to implement but has the potential to add tremendous value to investment screening and due diligence. There are three key contributions in this research proposal: Firstly, this proposal aims to validate the investment decision criteria proposed by surveyed PE managers gathered by Block et al (2019). Secondly, support arguments on PitchBook being a suitable database for future empirical research, especially with intersections between data science and private equity. Lastly, provide evidence to PE managers data science can inform investment due diligence and create efficiencies in screening for investments.

7.2 Future Research

There are several new avenues for research depending on a successive outcome(s) from this proposal. Firstly, this proposal only considers relative performance of investment criteria from the perspective of the average PE investor. Further research could explore the segmentation of PE investor types (family office, business angel, venture capital, growth equity, leveraged buyout). Lastly, this proposal could explore the implications of screening different industry segmentations and their alignment with different fund mandates.

7.3 Research Timetable

The implementation of this proposal will take place over a 14 week window, starting the July 19th 2021 and ending October 22nd 2021. The timetable in 2 in the Appendix (8.5) outlines the time taken to implement each step of the methodology outlined in Section 6. Additionally, the research timetable includes expectations on time commitments to report writing, reviewing, editing and meetings with supervisors.

References

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Management Journal*, *22*(6-7), 493–520. cited By 2567. doi:10.1002/smj.187

Ang, A., Chen, B., Goetzmann, W. N., & Phalippou, L. (2018). Estimating private equity returns from limited partner cash flows. *The Journal of Finance*, *73*(4), 1751–1783.

Aurélien, G. (2017). Hands-on machine learning with scikit-learn & tensorflow. *Geron Aurelien*.

Axelson, U., Jenkinson, T., Strömberg, P., & Weisbach, M. S. (2013). Borrow cheap, buy high? the determinants of leverage and pricing in buyouts. *The Journal of Finance*, *68*(6), 2223–2267.

BCG. (2019). Creating value in private equity with advanced data and analytics. Available at https://www.bcg.com/industries/principal-investors-private-equity/creating-value-private-equity-advanced-data-analytics (2021/06/05).

Bengtsson, O., & Sensoy, B. A. (2011). Investor abilities and financial contracting: Evidence from venture capital. *Journal of Financial Intermediation*, *20*(4), 477–502.

Bernstein, S., Giroud, X., & Townsend, R. R. (2016). The impact of venture capital monitoring. *The Journal of Finance*, *71*(4), 1591–1622.

Bernstein, S., Korteweg, A., & Laws, K. (2017). Attracting early-stage investors: Evidence from a randomized field experiment. *The Journal of Finance*, *72*(2), 509–538.

Block, J., Fisch, C., Vismara, S., & Andres, R. (2019). Private equity investment criteria: An experimental conjoint analysis of venture capital, business angels, and family offices. *Journal of Corporate Finance*, *58*, 329–352. doi:https://doi.org/10.1016/j.jcorpfin.2019.05.009

Bloomberg. (2020). Blackstone's next product may be data from companies it buys. Available at https://www.bloomberg.com/news/articles/2020-12-15/blackstone-s-next-product-could-be-data-from-companies-it-buys (2021/06/06).

Bottazzi, L., Da Rin, M., & Hellmann, T. (2008). Who are the active investors?: Evidence from venture capital. *Journal of Financial Economics*, *89*(3), 488–512. doi:https://doi.org/10.1016/j.jfineco.2007.09.003

Braun, R., Jenkinson, T., & Stoff, I. (2017). How persistent is private equity performance? evidence from deal-level data. *Journal of Financial Economics*, *123*(2), 273–291.

Brav, A., Jiang, W., Partnoy, F., & Thomas, R. (2008). Hedge fund activism, corporate governance, and firm performance. *The Journal of Finance*, *63*(4), 1729–1775.

Brown, G. W., Harris, R. S., Jenkinson, T., Kaplan, S. N., & Robinson, D. T. (2015). What do different commercial data sets tell us about private equity performance? *Available at SSRN 2706556*.

Cavagnaro, D. R., Sensoy, B. A., Wang, Y., & Weisbach, M. S. (2019). Measuring institutional investors' skill at making private equity investments. *The Journal of Finance*, *74*(6), 3089–3134.

Chemmanur, T. J., Krishnan, K., & Nandy, D. K. (2011). How does venture capital financing improve efficiency in private firms? a look beneath the surface. *The Review of Financial Studies*, *24*(12), 4037–4090.

Crunchbase. (2021). Crunchbase. Available at https://www.crunchbase.com/ (2021/06/06).

Curtin, K. M. (2018). 1.12 - network analysis. In B. Huang (Ed.), *Comprehensive geographic information systems* (pp. 153–161). doi:https://doi.org/10.1016/B978-0-12-409548-9.09599-3

Edgar, T. W., & Manz, D. O. (2017). Chapter 4 - exploratory study. In T. W. Edgar & D. O. Manz (Eds.), *Research methods for cyber security* (pp. 95–130). doi:https://doi.org/10.1016/B978-0-12-805349-2.00004-2

Ewens, M., & Rhodes-Kropf, M. (2015). Is a vc partnership greater than the sum of its partners? *The Journal of Finance*, *70*(3), 1081–1113.

Faudree, R. (2003). Graph theory. In R. A. Meyers (Ed.), *Encyclopedia of physical science and technology (third edition)* (Third Edition, pp. 15–31). doi:https://doi.org/10.1016/B0-12-227410-5/00296-9

Franke, N., Gruber, M., Harhoff, D., & Henkel, J. (2008). Venture capitalists' evaluations of start–up teams: Trade–offs, knock–out criteria, and the impact of vc experience. *Entrepreneurship Theory and Practice*, *32*(3), 459–483. doi:10.1111/j.1540-6520.2008.00236.x. eprint: https://doi.org/10.1111/j.1540-6520.2008.00236.x

Gompers, P., Kaplan, S. N., & Mukharlyamov, V. (2016). What do private equity firms say they do? *Journal of Financial Economics*, *121*(3), 449–476.

Gompers, P., & Lerner, J. [Josh]. (2001). The venture capital revolution. *Journal of economic perspectives*, *15*(2), 145–168.

Harris, R. S., Jenkinson, T., & Kaplan, S. N. (2014). Private equity performance: What do we know? *The Journal of Finance*, *69*(5), 1851–1882.

Hellmann, T., & Puri, M. (2002). Venture capital and the professionalization of start-up firms: Empirical evidence. *The journal of finance*, *57*(1), 169–197.

Hellmann, T., Schure, P., & Vo, D. (2013). Angels and venture capitalists: Complements or substitutes? *NBER Working paper*.

IBM. (2021a). Natural language processing (nlp). Available at https://www.ibm.com/cloud/learn/natural-language-processing (2021/06/06).

IBM. (2021b). What is data science? Available at https://www.ibm.com/cloud/learn/data-science-introduction (2021/06/05).

Iliev, P., & Lowry, M. (2020). Venturing beyond the ipo: Financing of newly public firms by venture capitalists. *The Journal of Finance*, *75*(3), 1527–1577.

Insights, C. (2021). Cb insights. Available at https://www.cbinsights.com/ (2021/06/06).

Kaplan, S. (1989). The effects of management buyouts on operating performance and value. *Journal of financial economics*, *24*(2), 217–254.

Kaplan, S. N., & Lerner, J. [Josh]. (2016). Venture capital data: Opportunities and challenges. *Measuring entrepreneurial businesses: current knowledge and challenges*, 413–431.

Kaplan, S. N., & Schoar, A. (2005). Private equity performance: Returns, persistence, and capital flows. *The journal of finance*, *60*(4), 1791–1823.

Kaplan, S. N., & Sensoy, B. A. (2015). Private equity performance: A survey. *Annual Review of Financial Economics*, *7*, 597–614.

Kaplan, S. N., & Stromberg, P. (2001). Venture capitals as principals: Contracting, screening, and monitoring. *American Economic Review*, *91*(2), 426–430.

Kaplan, S. N., & Stromberg, P. (2009). Leveraged buyouts and private equity. *Journal of economic perspectives*, *23*(1), 121–46.

Kaplan, S. N., & Strömberg, P. E. (2004). Characteristics, contracts, and actions: Evidence from venture capitalist analyses. *The Journal of Finance*, *59*(5), 2177–2210.

Korteweg, A., & Nagel, S. (2016). Risk-adjusting the returns to venture capital. *The Journal of Finance*, *71*(3), 1437–1470.

Korteweg, A., & Sorensen, M. (2017). Skill and luck in private equity performance. *Journal of Financial Economics*, *124*(3), 535–562.

Lerner, J., Sorensen, M., & Strömberg, P. (2011). The long-run impact of private equity: The impact on innovation. *Journal of Finance*, *66*, 445–78.

Lerner, J. [Josh]. (1995). Venture capitalists and the oversight of private firms. *the Journal of Finance*, *50*(1), 301–318.

Lerner, J. [Josh], Schoar, A., & Wongsunwai, W. (2007). Smart institutions, foolish choices: The limited partner performance puzzle. *The Journal of Finance*, *62*(2), 731–764.

MergerMarket. (2021). Mergermarket. Available at https://www.mergermarket.com/info/ (2021/06/06).

Paglia, J. K., & Harjoto, M. A. (2014). The effects of private equity and venture capital on sales and employment growth in small and medium-sized businesses. *Journal of Banking & Finance*, *47*, 177–197. doi:https://doi.org/10.1016/j.jbankfin.2014.06.023

Phalippou, L. (2020). An inconvenient fact: Private equity returns and the billionaire factory. *The Journal of Investing*, *30*(1), 11–39.

Phalippou, L., & Gottschalg, O. (2009). The performance of private equity funds. *The Review of Financial Studies*, *22*(4), 1747–1776.

PitchBook. (2021). Pitchbook. Available at https://pitchbook.com/ (2021/06/06).

Preqin. (2021). Preqin. Available at https://www.preqin.com/ (2021/06/06).

Puri, M., & Zarutskie, R. (2012). On the life cycle dynamics of venture-capital- and non-venture-capital-financed firms. *The Journal of Finance*, *67*(6), 2247–2293. doi:https:

//doi.org/10.1111/j.1540-6261.2012.01786.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2012.01786.x

PYPL. (2021). Pypl popularity of programming language. Available at http://pypl.github.io/PYPL.html (2021/06/13).

Rin, M. D., Hellmann, T., & Puri, M. (2013). Chapter 8 - a survey of venture capital research. In G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.), (Vol. 2, pp. 573–648). Handbook of the Economics of Finance. doi:https://doi.org/10.1016/B978-0-44-453594-8.00008-2

Robinson, D. T., & Sensoy, B. A. (2013). Do private equity fund managers earn their fees? compensation, ownership, and cash flow performance. *The Review of Financial Studies*, *26*(11), 2760–2797.

Ross, G., Das, S., Sciro, D., & Raza, H. (2021). Capitalvx: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, A., & Teal, T. (2016). Good enough practices in scientific computing. *PLOS Computational Biology*, *13*. doi:10.1371/journal.pcbi.1005510

8 Appendix

8.1 Mathematics

This subsection informs the mathematical expressions pertaining to data preparation and algorithm implementation.

8.1.1 Performance Measures

Root Mean Square Error

$$RMSE(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x}^{(i)}) - y^{(i)})^2} \tag{3}$$

$$\tag{4}$$

Mean Absolute Error

$$MAE(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}|(h(\mathbf{x}^{(i)}) - y^{(i)})| \tag{5}$$

X is a matrix containing all feature values, of all instances, from the dataset.

8.1.2 Multi-level Logistic Regression

Generalised Multi-level Logistic Regression

$$ln(\frac{\hat{p}}{1-\hat{p}}) = \beta \cdot X + \mathcal{E} \tag{6}$$

- $\hat{p}$: The expected probability that the outcome is present.
- β: The vector of co-efficient related to sensitivities
- X: The vector of distinct independent variables.
- $\mathcal{E}$: The vector of error terms.

8.1.3 Multi-nominal Logistic Regression (Softmax Regression)

Softmax Score for Class K

$$s_k(\mathbf{x}) = \theta_k^T \cdot \mathbf{x} \tag{7}$$

$$\tag{8}$$

Softmax Function

$$\hat{p}_k = \frac{exp(s_k(\mathbf{x}))}{\sum_{j=1}^{K} exp(s_k(\mathbf{x}))} \tag{9}$$

$$\tag{10}$$

Softmax Regression Classifier Prediction

$$\hat{y}_{\cdot()} = \text{argmax}_k \ \sigma(s_k(\mathbf{x})) = \text{argmax}_k \ \sigma(\theta_k^T \cdot \mathbf{x}) \tag{11}$$

$$\tag{12}$$

Cross Entrophy Cost Function

$$J(\Theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y_k^{(i)} log(\hat{p}_k^{(i)}) \tag{13}$$

$$\tag{14}$$

Cross Entropy Gradient Vector for Class K

$$\nabla_{\theta_k} J(\Theta) = \frac{1}{m} \sum_{i=1}^{m} (\hat{p}_k^{(i)} - y_k^{(i)}) \mathbf{x}^{(i)} \tag{15}$$

$$\tag{16}$$

There are two sets of classes: k_1 ϵ\{Investment (1,0)\} and k_2 ϵ\{IPO, Acquistion, Bankrutpcy (1,0)\}. $s_k(\mathbf{x})$ is the score for each class k. $\hat{p}_k$ that the instance belongs to class k. θ_k is the parameter vector for class k. Θ is the parameter matrix containing all parameter vectors.

K is the number of classes. $s(\mathbf{x})$ is the vector containing all scores of each class for the instance $\mathbf{x}$. $\sigma(s(\mathbf{x}))_k$ is the estimated probability probability that the instance $\mathbf{x}$ belongs to class k given the scores of each class for that class. Argmax_k returns the value of k that maximises the estimated probability of $\sigma(s(\mathbf{x}))_k$. $y_k^{(i)} = 1$ if the target class for the ith instance is k, 0 otherwise.

8.1.4 Random Forests

Gini Impurity

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{17}$$

$$\tag{18}$$

Entropy

$$H_i = 1 - \sum_{k=1, P_{i,k} \neq 0}^{n} p_{i,k} log(p_{i,k}) \tag{19}$$

$$\tag{20}$$

CART Cost Function (Gini)

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right} \tag{21}$$

$$\tag{22}$$

CART Cost Function (Entropy)

$$J(k, t_k) = \frac{m_{left}}{m} H_{left} + \frac{m_{right}}{m} H_{right} \tag{23}$$

$$\tag{24}$$

Variables

- $p_{i,k}$ is the ratio of class k instances among the training instances in the i^{th} node.

- $G/H_{left/right}$ measures the impurity of the left.right subset using gini or entropy measure respectively.
- $m_{left/right}$ is the number of instances in the left/right subset.

8.1.5 Multi Layer Perceptron (MLP)

Perceptron Learning Rule

$$w_{i,j}^{\text{next step}} = w_{i,j} + \eta(\hat{y}_j - y_j)x_i \tag{25}$$

Variables

- $w_{i,j}$ is the connection weights between the ith input neuron and the jth output neuron.
- x_i is the ith input value of the current training instance.
- $\hat{y}_j$ is the output of the jth output neuron for the current training instance.
- y_j is the output of the jth output neuron for the current training instance.
- η is the rate.

8.2 Technology

The subsequent technologies enable the creation a project workspace and the implementation of the research methodology.

- Python: open-source, interpreted programming language
  - **Numpy**: large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
  - **Pandas**: data analysis and manipulation functionalities.
  - **Scikit-Learn**: software machine learning library for the Python programming language.

  - **TensorFlow**: open-source software library for machine learning with a focus on training and inference of deep neural nets.
  - **Matplotlib**: plotting functionalities ported from MatLab.
  - Additional libraries when required.
- **Anaconda**: distribution service of the Python and R programming languages
- **macOS Mojave (OS)**: operating system.
- **MacBook Pro 2019**: hardware.
  - 13inch
  - 1.4 GHz Intel Core i5
  - 8 GB 2133 MHz LPDDR3
- **Microsoft Visual Studio Code**: integrated programming environment (IDE).
- **Git & GitHub**: version control.
- **IBM Watson**: suite of AI-related products and functionalities
- **Amazon Web Services**: provides cloud computing technologies, platforms and APIs.

8.3 Case Studies

A couple of case studies inform the the practicality of data science in PE. BCG (2019) published analysis on creating value in Private Equity with Advanced Data and Analytics, providing three key examples: geo-analytics, predictive maintenance, and workforce optimisation. Geo-analytics identifies profitable locations for vending machines. Predictive maintenance prioritizes repairing machines with higher failure risks to mitigate repair and replacement costs. Workforce optimisation matches the skillsets of technicians with customer requirements. Blackstone employ data scientists to inform both portfolio operations and investment practices. The buyout fund has the unique ability to sell

data of portfolio companies and create value for the owner (Bloomberg, 2020). From an investing perspective, NZ Super Fund are exploring data science applications in equities selection methods to compare against traditionally equity selection processes. The comments on BCG and Blackstone inform operational applications of data science while the comments on NZ Super Fund inform investment applications. These case studies frame inform applications on how data science can add value to PE.

8.4 Investor Criteria, Attributes & Variables

Table 6
Attributes and attribute levels used in our conjoint analysis.
This table describes and defines the attributes and attribute levels presented to participants in our conjoint analysis. We use a choice-based-conjoint (CBC) analysis, in which the participants are presented with investment opportunities and are asked to select the one company that better matches their preferences. The two companies are only described in terms of the attributes displayed in this table ("attributes") and only differ from each other in the respective specification of these criteria ("attribute levels").

Attribute	Attribute levels	Attribute description
(1) Profitability (3 levels, ordinal)	1. *Not profitable* 2. *Break-even* 3. *Profitable*	Describes the current profitability of the company.
(2) Revenue growth (4 levels, ordinal)	1. *10% p.a.* 2. *20% p.a.* 3. *50% p.a.* 4. *100% p.a.*	Represents the company's average yearly revenue growth rate over the last years.
(3) Track record management team (3 levels, ordinal)	1. *None of them* 2. *Some of them* 3. *All of them*	Describes whether the management team has a relevant track record (e.g., industry experience or leadership experience).
(4) Current investors (3 levels, nominal)	1. *No other current external investors* 2. *Other current external investor - Unfamiliar* 3. *Other current external investor - Tier 1*	Describes the type of current investor, if any.
(5) Business model (4 levels, nominal)	1. *Lock-in* 2. *Innovation-centered* 3. *Low cost* 4. *Complementary offering*	Describes the key focus of the business model of the company: 1. Lock-in: Business model that keeps customers attracted and "locked-in", having high switching costs for customers, which prevent them from changing to other providers. 2. Innovation-centered: Business model that offers innovation in the form of new technology, products or services. 3. Low cost: Business model focusing on reducing costs for customers for already existing products or services. 4. Complementary offering: Business model that bundles multiple goods or services to generate more value for customers.
(6) Value-added of product/ service (3 levels, ordinal)	1. *Low* 2. *Medium* 3. *High*	Describes the value added for the customer through the product or service. Low value-added represents a marginal improvement (e.g., in cost-reduction or service quality), whereas high value-added represents significant improvements.
(7) International scalability (3 levels, ordinal)	1. *Easy* 2. *Moderate* 3. *Difficult*	Describes the difficulty of scaling the company internationally, in terms of the time and investment needed.

Figure 1: Investment criteria and attributes from Block et al (2019)

Criteria	Relative Importance (Attributes,%)	Rank (#)	Variables
Revenue Growth	23.4	1	Revenue Growth (t-1, t-3, t-5) 10%, 20%, 50%, 100% (p.a)
Value-added (Product/Service)	20.4	2	Low (1), Med (2), High (3)
Management Team Track Record	15.7	3	None (1), Some (2) , All (3)
International Scalability	13.0	4	Easy (1), Moderate (2), Difficult (3)
Profitability	11.8	5	EBITDA, EBIT, NOPAT, ROA, ROE ($) Not profitable (1), Breakeven (2), Profitable (3)
Business Model	8.3	6	Lock in (1), Innovation-centered (2), Low Cost (3), Complimentary Offering (4)
Current Investors	7.3	7	No Other Current External Investors (1), Other Current Investors - Unfamiliar (2) Other Current Investors - Tier 1 (3)

Table 1: Variables mapping investment criteria

8.5 Research Timetable

This section of the appendix displays the proposed timeline for the research proposal. The timetable displays the implementation of the proposal's methodology through time.

Tasks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Comments
	19/07/21	26/07/21	2/08/21	9/08/21	16/08/21	23/08/21	30/08/21	6/09/21	13/09/21	20/09/21	27/09/21	4/10/21	11/10/21	18/10/21	First day of the week
Problem Scoping															Review to ensure scoping is correct
Data Acquisition															Contact industry connections to arrange access to PitchBook
Discovery & Visualisation															Outlined in research proposal
Data Preparation															Outlined in research proposal
Model Selection & Training															Outlined in research proposal
Model Tuning															Outlined in research proposal
Solution Presentation															Outlined in research proposal (dissertation writing separate)
Report Writing															Make weekly contributions, writing every day as step through methodology
Review/Editing															Review dissertation deliverable between initial and final submission dates
Submission (22/10/2021)															Split into an preliminary submission for feedback (4/10/21), receive from supervisor to edit (11/10/21) and final submission (22/10/21)
Supervisor meeting															Meet with supervisor once every two weeks to discuss progress, dissertation problems and resolutions

Figure 2: Research Timetable