Handelshøyskolen BI

GRA 19703 Master Thesis

Thesis Master of Science 100% - W

Predefinert informasjon

Startdato:	16-01-2022 09:00	**Termin:**	202210
Sluttdato:	01-07-2022 12:00	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	T		
Flowkode:	202210||10936||IN00||W||T		
Intern sensor:	(Anonymisert)		

Deltaker

Navn:	Jurij Starman

Informasjon fra deltaker

Tittel *:	Predicting private equity fund returns
Navn på veileder *:	Tatyana Marchuk

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	201
Andre medlemmer i gruppen:	Deltakeren har innlevert i en enkeltmannsgruppe

Predicting Private Equity Fund Returns

Master Thesis

Jurij Starman

MSc in Finance

Supervisor: Tatyana Marchuk

Oslo, July 1, 2022

ABSTRACT

This thesis investigates the potential of a Private Equity fund performance forecasting model, to assist Private Equity investors in their investment decision making process. Fund performance is measured by the fund's Kaplan Schoar Public Market Equivalent and is forecasted using a binary classification approach. The top performing Machine Learning models are able to forecast Buyout fund performance with 63 % accuracy, and Venture Capital fund performance with 66 % accuracy. Therefore, the features used to train the models and selected based on the literature on Private Equity performance drivers, possess important predictive power, which can be integrated in the investment procedure.

Keywords: Private Equity, Machine Learning, Public Market Equivalent, Buyout, Venture Capital, Financial Forecasting

Acknowledgements

I would like to thank my supervisor Tatyana Marchuk for assistance and advice.

Contents

List of Abbreviations

The following table describes the meaning of various abbreviations and acronyms used throughout the thesis. The page, on which each one is defined or first used, is also provided.

Abbreviation	Meaning	Page
PE	Private Equity	1
USD	United States Dollars	1
AUM	Assets Under Management	1
ML	Machine Learning	1
AI	Artificial Intelligence	1
GDP	Gross Domestic Product	1
(KS-)PME	(Kaplan Schoar) Public Market Equivalent	2
LR	Logistic Regression	2
kNN	k-Nearest Neighbours	2
SVC	Support Vector Classifier	2
SVM	Support Vector Machine	2
DTC	Decision Tree Classifier	2
RFC	Random Forest Classifier	2
BO	Buyout	2
VC	Venture Capital	2
IRR	Internal Rate of Return	3
GP	General Partner	9
LP	Limited Partner	9
MBO	Management Buyout	11
MBI	Management Buy-In	11
LBO	Leveraged Buyout	11
P2P	Public-To-Private	11
PIPE	Private Investment to Public Equity	11
DPI	Distributed to Paid-In	12
RVPI	Residual Value to Paid-In	12
TVPI	Total Value to Paid-In	12
CV	Cross Validation	16
NA	North America	17
NAV	Net Asset Value	20
ROC	Receiver Operating Characteristic	34
AUC	Area Under the Curve	34
TPR	True Positive Rate	34
FPR	False Positive Rate	34

List of Figures

List of Tables

1 Introduction

Since its inception in 1946 the Private Equity (PE) Market has provided imperative funding to start-up companies, private companies, companies in financial distress and companies seeking buyout financing. The market's rapid growth and increased importance, which started in the 1980s, has continued throughout the turn of the century, reaching 8 trillion USD AUM in 2021 (Preqin, 2022) – four times as much as in 2010. This expansion has been accompanied by an increase in the number of funds, with a significant performance gap between top and bottom quartile funds, which reached a mean of 13.15% from 2000 to 2016 (Preqin, 2022). As such, the difficulty, and the cost of the fund selection process for PE investors has increased. Concurrently, the use of Machine Learning (ML) in finance has expanded. However, it was mostly limited to areas such as cross-sectional stock market prediction, bankruptcy prediction, and default recovery rates. PE was relatively slow in incorporating the newly available digital tools. Currently, ML use in PE is mostly limited to PE firms, which employ Artificial Intelligence (AI), data mining, and web-based analytics to assist in the investment company selection process (Bain's PE report, 2022). Consequently, the question arises about the viability of ML techniques to assist investors in their PE investment decision making process. Current research of ML application in PE is severely limited. However, since the traditional approach that PE investors use to select promising investments is based on a set of criteria (e.g., fund-level statistics, past fund performance), these 'performance drivers' have been extensively investigated. Moreover, research on ML applications in other areas of finance (e.g., Gu et. al., 2019) is plentiful. Therefore, I draw from the findings of these research areas to investigate the question: **Can ML tools be used to assist in the PE investors fund selection process?**

I conduct the investigation using a sample of 1434 funds, with vintages ranging from 1985 to 2017. To train the ML models I use fund-level statistics (e.g., fund size) and macroeconomic data (e.g., GDP), selected based on prior literature and data availability. I translate the fund selection process into a binary classification problem, where the classification is based on whether a fund is predicted to outperform or underperform a selected performance benchmark. The metric

selected to measure performance in this thesis, is the Kaplan Schoar Public Market Equivalent (KS-PME) measure, while the models used to predict fund performance are: Logistic Regression (LR), k-Nearest Neighbours (kNN), Support Vector Classifier (SVC), Support Vector Machine (SVM), Decision Tree Classifier (DTC), and Random Forest Classifier (RFC). The models' performances are subsequently compared and a possible explanation behind the differences in their performance is given. Furthermore, the LR and RFC models offer insight into the contribution of different performance drivers to the classification process. Consequently, I discuss the findings of the feature importances and relate them to the existing literature.

The analysis yields promising results in terms of ML use in the investment decision making process with the SVC performing best for the Buyout (BO) dataset and DTC and SVM performing best for the Venture Capital (VC) dataset. Moreover, all the models outperform the random classifier. While the analysis can be improved in many ways it still demonstrates the viability of ML use in the PE investor's investment decision making process. Thus, it can serve as an incentive for further research in the application of ML in PE asset space.

The subsequent parts of the thesis are organized as follows. Section 2 provides an overview of the existing literature on PE performance, performance drivers, and ML in finance. In Section 3, I describe the necessary theoretical characteristics of PE and ML. Section 4 includes a description of the PE funds used in the sample and the feature engineering required to train the ML models. Section 5 describes the methodology behind the approach to binary classification problems of the selected ML algorithms, as well as the measures, which are used to compare their performance. In Section 6 the results of the analysis are provided and discussed. The conclusive Sector 7 summarizes the main findings of the thesis and discusses the limitations and potential improvements which can be applied to the analysis.

2 Literature Review

Machine Learning has been rapidly integrated into various areas of finance, due to an increase in the availability of data and the reduction in the cost of computing power. This thesis researches the possible application of ML techniques in the PE investment decision making process. Due to the relative scarcity of the data available to PE investors, the research regarding this specific topic is somewhat limited. Consequently, I will discuss two main strands of literature, which relate to the research area of this thesis: PE performance and performance drivers, and Machine Learning application in finance.

2.1 Private Equity Performance and Performance Drivers

2.1.1 Performance

The traditional and still the most frequently used measures of PE performance are the Internal Rate of Return (IRR) and the money multiple (Gompers et. al., 2016). Compound returns have been chosen as the most appropriate performance metric over traditional annual return, due to the uncertain timing and amount of cash flows of a PE fund (Fraser-Sampson, 2010). Using a combination of money multiples and the IRR mitigates some of the well-known limitations associated with using only the IRR (it does not always exist, there can be multiple, it can be very sensitive to moving the timing of cash flows etc.) and using only the multiples (they do not account for the timing of cash flows). However, the most important drawback of using IRR and multiples as performance metrics is that they do not account for the risk associated with the investment. Moreover, Phallipou and Gottschalg (2009) highlight other potential problems regarding IRR in PE, most notably the significant upward bias of average IRRs.

Long and Nickels (1996) were the first to introduce a new kind of performance metric called the PME, which was popularized and redefined by Kaplan and Schoar (2005). The PME performance measurement relates an investment in a PE fund to an investment in a public equity index (originally and most commonly to the S&P 500). Since its origin there have been many improvements and generalizations of the seminal KS-PME, most notably by Korteweg and Nagel (2013), who relax the assumptions of the traditional PME by introducing an adapted SDF method.

However, the KS-PME is still the most commonly used PME metric (eVestment, 2017) and Sorensen and Jagannathan (2013) provide a thorough explanation as to why it is still valid. Furthermore, recognizing the value of the PME performance metric, PE investors have started increasingly integrating it, in addition to the standard performance metrics, in their due diligence process (eVestment, 2017). However, the survey conducted by Gompers et. al. (2016), shows that by far the most commonly used metric is still the multiple of invested capital, followed by the IRR. The paper by Harris et. al. (2014b) demonstrates that all three main performance metrics are correlated, with IRRs and money multiples reliably predicting PMEs.

To conclude, due to the nature of PE fund cash flows, traditional methods of performance measurement are not applicable. Consequently, the combination of IRR and money multiple is used, which has significant and well-known drawbacks, most notably they do not account for risk. The PME metric mitigates this limitation. While the money multiple and IRR are still the predominant metrics used in the industry, PME is becoming increasingly popular.

2.1.2 Performance Drivers

There have been several drivers of PE performance investigated in the literature. Fenn et. al. (2001) investigated the effect of aggregate amount of committed capital on PE returns. They concluded that the partnerships, which were formed during periods when small amounts of capital were raised, exhibited relatively high returns, while funds formed during periods when large amounts of capital were raised, exhibited low returns. They argue that the reason for this is the breakdown in discipline in deal pricing and structuring during times of greater capital availability.

Kaplan and Schoar (2005) further investigate the impact of capital flows on performance. Moreover, they extend the base of potential performance indicators to fund size, persistence, and overall fund manager survival. Their findings confirm the discovery of Fenn et. al. (2001) that funds raised in boom times of the PE industry perform poorly, however they find that the poor performance is mainly driven by new entrants into the industry, and that the performance of more established funds is less affected. They attribute this disparity to the heterogeneity in the skill and quality of fund managers. Furthermore, they find that larger funds, managed by more experienced managers perform significantly better. Additionally,

they observe a concave relationship between fund size and performance i.e., larger funds perform better, but when funds become very large, performance declines, and a convex relationship between manager experience and performance. However, this finding was not as significant as for fund size. Moreover, they detect significant persistence in fund returns across different funds, managed by the same manager.

Similarly, to Kaplan and Schoar (2005), Phallippou and Zollo (2005) find that low performance is concentrated in small and inexperienced funds. In addition, they find that PE fund performance is significantly procyclical, that it increases with the average GDP growth rate and decreases with the average level of interest rates (both measured by multiple different proxies). Furthermore, they observe that performance increases with the average return on the stock market index (CRSP - VW index).

Lossen (2006) expands the investigation of performance predictors by examining the effects of diversification across financing stages, industries, and countries. The outstanding theoretical hypothesis states that due to significant information asymmetry in the industry, the expected outcome of diversification is that it harms returns. However, the author discovers that the rate of return of PE funds does indeed decline with diversification across financing stages, however, it increases with diversification among industries, and is not affected by diversification across countries. Additionally, he finds a strong negative link between rate of return of the MSCI World Index in vintage year, which is consistent with the findings of Kaserer and Diller (2009). He credits this negative relation to PE firms having to pay high prices for their investments when the global economy is performing well. The author also detects a discrepancy, compared to other literature, between the relation of fund size and the amount of new funds raised by the global PE industry, to fund performance. He detects a decrease in returns with the increase of fund size and an increase in returns with an increase in the number of new entrants.

Aigner et. al. (2008), observe the same discrepancy with regards to fund size, however they provide an explanation, stating that using a squared term in the regression (similarly to Kaplan and Schoar (2005)) might result in a sign switch. Additionally, Aigner et. al. (2008) provide an overview of the impact of both endogenous (region, industry sector, financing stage, vintage year and fund manager) and exogenous (performance of the public market, interest rates, GDP growth) factors on fund performance. In accordance with other literature, they find

that fund manager experience, GDP growth and average MSCI growth is positively related to performance, and that interest rates yield a negative influence on performance. They argue that in times when interest rates are high, the cost of financing increases, resulting in lower returns. Interestingly, their investigation yielded a negative influence of public equity market growth (MSCI) to fund performance, despite a positive influence of average MSCI growth. Contrary to Lossen (2006), they found no significant impact of industry diversification and a positive relation of diversification across financing stages. Furthermore, they have corroborated Lossen's (2006) findings that country diversification does not impact fund performance.

More recent studies conducted by Roggi et. al. (2019) and Harris et. al. (2022) confirm the findings of Kaplan and Schoar (2005) that fund size and manager experience have a concave and convex relationship with fund performance, respectively. Moreover, they observe strong persistence in VC funds and a weakened persistence in BO funds, confirming that persistence has persisted in the PE industry. Additionally, Harris et. al. (2022) investigate persistence using information available to investors at the time of fundraising (rather than final fund performance). They find strong persistence for VC funds, but little evidence of persistence for BO funds.

To conclude, there have been many different PE fund performance drivers studied in the literature. They can be broadly separated into fund specific drivers and macroeconomic drivers. Fund specific drivers include the fund size, management experience, fund strategy (as in Buyout or Venture Capital) and diversification across financing stages, industries and countries/regions. Macroeconomic drivers include the average GDP growth rate, the average interest rate and the (average) return in a selected global public equity index (all sampled in the vintage year of the investigated funds). Additionally, PE industry specific factors can be included in the macroeconomic driver category. They include the aggregate amount of committed capital flowing into the PE industry and the number of new funds raised in a given year. The literature yielded mixed results in terms of the effect of the aforementioned performance drivers. However, if results from the newer studies are given higher validity, due to the greater availability of data to perform their investigations and the access to all the prior literature on which they improve upon, we can deduce that the results obtained in the early 2000s have held up surprisingly

well. Fund size and manager experience are shown to have a positive effect on performance, with the performance gains diminishing and even reducing in the case of fund size. Moreover, the average GDP growth rate and the average return on a global public equity index are more or less accepted to have a positive effect on fund performance. Additionally, the aggregate amount of capital flowing into the PE industry, the average interest rate, and the return on a public equity stock index are relatively universally accepted as having a negative effect on fund performance. The effects of the number of new funds entering the PE industry in a given year, and fund diversification across financing stages, industries, and countries have yielded contrasting results.

2.2 Machine Learning Application in Finance

Using ML techniques to questions in finance is not a particularly modern concept and its applications and capabilities have expanded in the recent decades. Back in the mid-1990s, Hutchinson et. al. (1994) used a so-called Learning Network (nonparametric method for performing nonlinear regressions) to price and hedge derivative securities. Lo et. al. (2002) used nonparametric kernel regression to deduce the validity of Technical Analysis. Gavrishchaka and Banerjee (2006) used SVM to forecast stock market volatility, and De Spiegeleer et. al. (2018) applied ML to accelerate derivative pricing.

The broad area of finance offers ML applications stemming from the classical SVM, kNN models (Farquad et. al., 2012; Imandoust and Bolandraftar, 2013), to modern DL techniques (Butaru et. al., 2016; Fischer and Krauss, 2018). ML application in finance has been mostly researched in relation to bankruptcy prediction (e.g., Zhao et. al., 2017), default recovery rates (e.g., Cheng et. al., 2018) and cross-sectional stock market prediction (e.g., Freyberger et. al., 2018). The expansion of studies in the last decade, is most likely because of the rising availability of financial data and the capability of ML techniques to process it efficiently and inexpensively (Warin and Stojkov, 2021).

In the field of PE, ML techniques have been adopted by PE firms (predominantly VC firms) to screen for favourable investment candidates. This is largely due to the high demand for PE financing and of course the abundance of proprietary data that these firms have acquired from the companies that they have previously invested, or plan to invest in. Using ML techniques to assist PE investors in their find

selection process has, however, been rarely researched in the academic circles. Largely due to PE being a relatively new asset class and therefore adequately large datasets have not been widely available, until recently. With the recent expansion of commercial PE data providers such as Preqin and Pitchbook, researching ML application will likely pose a lucrative challenge for researchers in the near future.

3 Theoretical Framework

3.1 Private Equity

Private Equity is one of the most misunderstood asset classes, largely due to it being a relatively young asset class and therefore evolving at a pace that exceeds the capabilities of linguistics experts to construct a formal definition. Consequently, the definitions vary significantly, ranging from 'A Private Equity investment is any equity investment in a company which is not quoted on a stock exchange' (Fraser-Sampson, 2010) to 'Private equity is the universe of all Venture and Buyout investing, whether such investments are made through funds, fund of funds or secondary investments' (EVCA, 2022). Regardless of the abundance of existing definitions, they rarely encompass all the aspects of the Private Equity asset class.

3.1.1 Private Equity as an Asset Class

The Private Equity asset class is categorized as being part of the alternative asset space. Alternative assets typically refer to investments that fall outside of the traditional asset classes, commonly accessed by most investors, such as stocks, bonds, or cash payments. They include but are not limited to Private Equity, Hedge Funds, Private Debt, Art and Antiques, Infrastructure, Natural Resources, and often, Real Estate. The traditional investments (meaning stocks, bonds, or cash) are traded via public markets and are subjected to heavy regulation from the financial regulatory authorities such as the SEC (Securities and Exchange Commission) or the FCA (Financial Conduct Authority). Contrarily, the alternative assets are traded privately and are often not heavily regulated. This lack of regulation often leads to alternative asset investments being available only to accredited investors. The reason accredited investors choose to invest in these alternative assets is because of their seemingly appealing risk-return characteristics. However, the private nature of these investments makes the assessment of their true risk-adjusted performance difficult, and therefore subject to debate. The measurement or assessment of the true risk-adjusted performance of these alternative asset classes surpasses the scope of this thesis.

Investments in Private Equity can either be made directly i.e., an investor directly buys the shares of private companies, or indirectly i.e., via a PE firm. The firm, also known as the General Partner (GP), raises funding from external investors, also known as Limited Partners (LPs). The LPs are passive, meaning they take no part

in the business and have limited liability. They are comprised of institutional and accredited investors, who seek returns which exceed the return of the public equity markets, by investing in alternative asset classes such as Private Equity. The GPs role and responsibility is to invest the fund's capital in investment companies, to actively manage the investments in the portfolio, seeking to generate operational improvements to increase the investment companies' value, and to seek to achieve exits with high returns. A fund will have a specific set of investment criteria although different funds within a firm may have different objectives. The firms themselves grow by raising new funds as existing funds approach maturity (continual process of fund raising and closing).

The mechanism of investing in a PE fund is different from investing in just about any other asset class. The investor does not invest all its committed capital i.e., the amount of capital which an investor has legally promised to provide to PE funds, at once. Instead, the capital is called when needed by the PE fund. Furthermore, when the PE fund sells an investment, the capital is distributed to LPs. This results in unpredictable cash flows coming in and out of the PE fund and result in the fund never actually holding money. Therefore, the fund acts as a conduit from the investments to the LPs. The firms typically invest around ninety percent of the total committed capital and reserve the remaining ten percent for additional investments in existing portfolio companies (used to cover operational costs, additional growth capital etc.).

The GPs remuneration is structured as follows. Firstly, the LPs are charged an annual management fee on committed capital, which usually amounts to around two percent. Secondly, the GPs acquire carried interest, which typically represents around twenty percent of the investment return generated above a minimum hurdle rate, which typically amounts to around seven to ten percent. The remaining eighty percent of the return on investment is distributed to LPs. Importantly, Private Equity investments are long-term investments. Typically, the fund has a life cycle of around ten years, out of which the first five to seven represent the so-called investment cycle i.e., the period in which the GP grows the investment portfolio and is therefore characterized by frequent capital calls and infrequent distributions. Additionally, the last three to five years represent the exit or harvesting cycle, which is, intuitively, characterized by numerous exits and consequently distributions to LPs.

Private Equity investments are most commonly categorized into Buyout and Venture (Capital). However, Growth (Capital) and Development (Capital) may be introduced as additional categories (Fraser-Sampson, 2010). The categories are selected based on where the investment company is in its life cycle. The firm's life cycle 'position' is closely related to the cash flows that it generates. Firms located in the 'early stage' generate negative cash flows, due to them having no product and/or service to sell and therefore no cash inflows. The firms in the 'growth' stage will have some inflows but its aggregate cash flows will still be strongly negative due to other costs e.g., promotion. In the 'maturity' stage, companies generate positive cash flows and are generally profitable. Lastly, in the 'decline' stage the companies generate, perhaps unintuitively, the highest cash flows (in theory, the reason for this is market consolidation). Importantly, the risk of a company surviving decreases, the further along its life cycle the company is located.

Venture Capital investments focus on firms in the 'early' and 'growth' stage. They can be further classified by Sector and Stage. The three main sectors in which VC investment firms are Life Science (often also referred to as BioTech or Healthcare), Information technology and Telecommunications. However, the distinction between the sectors is not absolute and has become increasingly blurred. Moreover, the main stages of VC investments are seed, early, mid, and late, which, intuitively, refer to the phase in which the VC firm has invested in the target company. As with the Sector classification, the distinction between the stages is not apparent, with Seed and Early stages frequently representing the source of confusion. The VC firm or fund generates its returns by purchasing shares in a (private) company, expecting to eventually be able to sell them for a higher price. Historically, the so-called 'home runs' (rare investment companies which have generated extraordinary returns) have driven the returns of VC firms, with less than 5% of companies by cost, generating 80% of the final fund value (Fraser-Sampson, 2010).

Buyout investments focus on firms in the 'maturity' and 'decline' stages. They are further categorized based on the different Buyout deals that occur in the PE market (see Table 1). However, the differences between the deal types are not clear-cut.

Buyout funds generate their returns, similarly to most PE investments, by selling the private company's equity for a higher price than it was purchased.

Table 1: Buyout strategy subcategories

The table represents the different Buyout deals that occur in the PE market as well as a brief description of each category (meanings of the category names are under Abbreviations).

Category	Description
MBO	Occurs when an executive or management team, who manages a particular business activity, decides to purchase said activity out of the parent company
MBI	Occurs when an executive or management team comes together to purchase another company, which operates in the same sector as the parent company
BIMBO	Occurs when outside executives are grafted on to the existing executive team in order to facilitate a buyout (combination of MBI and MBO)
LBO	Occurs when a buyout is not initiated by a management team (internal or external) but is instead initiated by a seller who appoints an investment bank to prepare a company for sale and then a buyout firm competes for ownership alongside industrial purchasers (the most blurred category since leverage is used in most buyout deals)
P2P	Occurs when a Buyout fund purchases a public company and de-lists it
Roll-up	Occurs when a Buyout fund purchases a lot of small operators in a fragmented industry and joins them
Secondary BO	Occurs when a Buyout fund's exit of a particular investment is not routed via an IPO or sold to a trade buyer but is instead sold to another Buyout firm or fund
PIPE	Occurs when a particular investment instrument is created in a public company that may offer a PE-type return and that company's equity is quoted but the instrument is not (the instrument is usually a convertible loan note with equity kickers)

However, Buyouts differ from the rest of the PE investment types in two important ways. Firstly, because the companies that they invest in already generate earnings

they can restructure the companies' capital to replace some of the equity with additional debt and are thus able to distribute returns to investors without exiting the deal, which in turn gives them the ability to take advantage of the time value of money and therefore generate higher IRRs. Secondly, they take advantage of tax consolidation i.e., the treatment of a firm which owns another firm as one large company for tax purposes. This gives them an ability to use large amounts of leverage, because they can take advantage of the tax shield on interest payments. Furthermore, they can use the investment company's own cash flows to repay the interest, which is also the main reason for targeting mature and/or declining companies. Historically, buyout deals have been increasing in size with as much as 95% of available buyout capital in Europe and North America targeting 5% of companies by number (Fraser-Sampson, 2010). This deal size increase has left a gap in the middle market (company value less than 500 Mn USD). Consequently, the gap has been filled by first time funds, in which many investors have a ban on investing. Moreover, because buyout deals use debt extensively, the loan terms which funds can negotiate are of paramount importance.

3.1.2 Private Equity Performance

Private Equity performance can be measured in various ways. The two categories of performance metrics are Absolute performance metrics and Market-adjusted performance metrics. The Absolute performance metrics include the IRR and investment multiples such as the Distributed to Paid-In (DPI), the Residual Value to Paid-In (RVPI), and the Total Value to Paid-In (TVPI). The Market-adjusted performance metrics are the PME, which compare an investment in a Private Equity fund to an investment in a selected market index. Furthermore, there exist several improvements of the PME, which are also categorized as Market-adjusted performance metrics, however they have not been widely adopted by investors and will therefore not be included as a performance measure in this thesis.

Compound returns (IRRs) have been universally accepted as the appropriate measure of performance in the PE industry. They measure the LPs annualized IRR based on fund contributions and distributions, net of fees and profit shares (carried interest). If a fund has not yet been liquidated, and therefore their final cash flow has not yet been revealed, the aggregate net values of the remaining assets of a fund are treated as that last 'cash-flow' in the calculation. The reason standard periodic returns cannot be used as a guide to PE performance is because of the uncertain

timing and amount of cash flows. The only certainty is that the total amount of cash inflows cannot exceed the amount of committed capital. Consequently, if we want the periodic returns to reflect the true return, we can calculate them only once the last cash flow has been paid out i.e., at the end of the fund's life.

If we want to observe the performance of a PE fund throughout the fund's lifetime we use the so-called ***J-curve***. The J-curve is produced by looking at the cumulative return of a fund to each year of its life (the first entry is the IRR of the fund's first year, the second entry the IRR for the first two years, the third for the first three years etc.). There exist differences in the shape of the J-curve, which depend on the fund's strategy and other investment specific factors. Typically, buyout funds tend to pay back their capital more quickly, which results in a quicker rising J-curve, whereas venture funds tend to pay back capital a bit later so the shape of the curve is flatter.

Multiples are a different way to look at PE fund performance and are typically used in tandem with the compound returns (IRRs). They are useful because they demonstrate the three-way relationship between the IRR, the multiple and the holding period. If a fund holds money for a longer period, it will have to deliver a higher multiple to sustain the same IRR. Consequently, the harder it is to 'put money to work' i.e., longer holding periods, the more multiples become a relevant measure of performance.

The ***DPI*** multiple compares the total amount of money paid out i.e., distributed to LPs to date, against the total amount of money paid into the fund by LPs. It is best used to measure performance of a fund once it is at the end of its life, because it shows the performance relative to all the money paid in, which includes fees and costs. It is not a good measure in two situations. Firstly, if the fund is not yet at the end of its life, because the fees and costs are high compared to invested capital. Secondly, if a fund has failed to invest all its capital, in which case fees and costs are again excessively high.

RVPI multiple shows the current value of all remaining investments i.e., companies within the fund. It is expressed as a ratio to the total amount paid-in to date. It is most useful as a measure early on in the life of a fund i.e., before there have been many distributions, because in that case it will reflect to what extent the portfolio companies may have been revalued. Its disadvantage is that it may give

misleadingly low return expectations because companies are typically sold for more than their current valuation (especially in the case of Buyout funds) (Brown et. al., 2019).

TVPI multiple is one of the most useful ratios. It adds together both the residual value and the distributions to date. Consequently, it is subject to the same possible drawbacks as RVPI.

A different approach to evaluating PE fund performance, which has gained traction in the recent years is the ***PME***. The PME compares an investment in a private equity fund to an investment in a selected market index. While many different versions of the PME exist, I will briefly describe the KS version, which will be used as a performance evaluator in the thesis. The KS-PME is implemented by discounting all cash outflows of the fund to the total return to the S&P500 and comparing the resulting value to the value of the cash inflows to the fund discounted using the total return to the S&P500. PME is a useful measure for LPs because it reflects the return to the private equity investment relative to public equities and, is the only performance metric, out of the aforementioned, that incorporates risk.

3.2 Machine Learning

ML was developed by answering the question 'how can computers learn to solve problems without being explicitly programmed' (Samuel, 1959). According to Mitchell (1997), the aim of ML is to produce systems whose performance improves with experience. To date ML has been applied to a wide range of problems such as data mining, game playing, speech and image recognition, as well as software and hardware testing (Bergadano and Gunetti, 1996). Considering the scope of the applicable areas, many approaches of solving these problems have been developed, stemming from a number of fields including genetics and statistics.

ML problems can be roughly categorized as supervised or unsupervised. In supervised learning the aim is to be able to predict an output measure based on one or multiple input measures. In unsupervised learning the aim is to discover some associations and/or patterns among a set of input measures. Additionally, supervised problems can be categorized into classification and regression. Regression is used when the goal is to predict quantitative outputs, while classification is used when the goal is to predict qualitative outputs. However, both

can be viewed as a task in function approximation. The focus of this thesis are supervised binary classification problems.

Furthermore, supervised problems can be solved in two distinct ways. Firstly, we can use the parametric method i.e., make an assumption about the form of the function we are trying to estimate. This allows us to select a suitable model, based on this assumption, and estimate a set of parameters in that selected model. The advantage of this approach is that the solution to the problem has very high interpretability, while the drawback is that the assumptions we made may not always hold. Secondly, we can use the non-parametric method, which, intuitively, does not make any (or very mild) underlying assumptions, with respect to the form of the function we are trying to estimate. The advantage of this approach is that it tends to be more accurate, while the drawback is that it is often less efficient.

The performance of parametric and non-parametric models is heavily dependent on the quantity of data we have available and/or the signal to noise ratio of that data. The less data we have or the 'noisier' the data is, the better parametric models tend to perform, compared to non-parametric models, and vice versa. This phenomenon occurs due to the so-called bias-variance trade-off. Mathematically, the trade-off can be represented as:

$$Err(x) = E\left[\left(Y - \hat{f}(x)\right)^2\right] \quad (1)$$

$$Err(x) = \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right] + {\sigma_e}^2 \quad (2)$$

$$Err(x) = Bias^2 + Variance + Irreducible\ Error$$

Bias is the estimation error between the actual value and the predicted value, which occurs due to generalization. Variance is the variability of the model prediction i.e., how much would our prediction change, if we estimated a model using a different data set. The trade-off occurs when we attempt to fit the data (approximate the function) in such a way that ensures the highest possible out-of-sample performance. If we underfit the data, the bias is too high. If we overfit the data, the variance is too high. Consequently, to ensure the best performance of the model we

strive to find the ‘correct’ amount of both bias and variance i.e., minimize the total error. In practice this trade-off is regulated by changing the hyper-parameters of the models (e.g., the λ in LR or the k in kNN).

Commonly, out-of-sample performance of the model is unknown. Therefore, we use techniques to estimate it and to ‘tune’ the hyper-parameters in a way that maximizes it. The most widely used technique is Cross Validation (CV). CV can be implemented in various ways. However, the idea behind all the implementations is similar: We separate the data into the training set and validation set, fit the model using the training set, and use the validation set to estimate the out of sample performance. We can use this approach just to estimate the out of sample performance of our model, to (continuously) change the hyper-parameters to achieve the maximal estimated out-of-sample performance, or to select the best performing model out of our selected model set.

4 Sample Selection and Data Description

4.1 Characteristics of PE Fund Data

In the recent years there has been significant growth in the amount of PE data accessible to researchers. This growth was spurred by the increase in the number of available commercial data suppliers. Currently the main commercial PE data providers are Burgiss, Cambridge Associates, Pitchbook, Preqin and Venture Economics. The reported data of the providers can be separated into performance related data (e.g., IRR, multiples), fund-level statistics (e.g., strategy, location, size) and cash flow level data (e.g., distributions, contributions).

The potential biases caused by the differences in reported data of the suppliers were investigated by Brown et. al. (2015). They concluded that for North American (NA) BO funds all reviewed data providers i.e., Cambridge Associates, Burgiss, PitchBook and Preqin, have similar sample sizes. However, there are some notable differences across databases in coverage of NA VC funds, which stem from the differences in data collection techniques employed by the data providers. Despite the differences in NA VC fund coverage, all reviewed data sources provide similar signals on fund performance for both NA VC and NA BO funds. Outside NA, coverage varies substantially across databases for BO funds. However, performance measures are relatively consistent. For VC funds outside NA both the coverage and performance vary significantly by database.

The PE related data used throughout this thesis has been sourced solely from the Preqin database. Consequently, the analysis could be improved by using a combination of data from different providers, which would minimize the selection bias. Preqin provides financial data and information on the alternative assets market, as well as tools to support investment in alternatives. Its data encompasses private capital and hedge funds, including fund, fund manager, investor, performance, and deal information. The asset classes it covers are PE, VC, hedge funds, private debt, real estate, infrastructure, natural resources, and secondaries. In addition to a variety of institutional investors Preqin collects performance data directly from fund managers. The performance figures from institutional investors are obtained via Freedom of Information Act (FOIA) requests (or their parallel outside the U.S.). Institutional LPs include CalPERS, Washington State Investment Board, and Florida State Board of Administration, among many others both in the

US and the UK. Additionally, fund managers of over 2200 firms submit a substantial proportion of Preqin's performance data, with Preqin reaching out to regular contributors every quarter to ensure the reported data is the latest available. Other sources of data include listed firm financial reports, public filings, and annual reports.

To ensure the collected data is consistent with Preqin's calculation methodologies, the GPs and FOIA (or their parallel outside the U.S.) sources must comply with certain guidelines when submitting their data. Moreover, Preqin has a designated internal Performance Team, who is tasked with reviewing the aforementioned data and cross-referencing it against a benchmark of similar funds, as well as against other sources reporting for the same fund. Consequently, the data provided on the database is as accurate as possible. The downside of the public approach to data collection is that the reliance on FOIA disclosures and voluntary submissions may lead to a sample that is not representative of the universe of funds. Because FOIA taps only certain types of investors (e.g., public pension funds) and because of voluntary submission (especially by GPs) the reported data may introduce selection and survivorship biases.

4.2 Sample Selection

To construct the sample of BO and VC funds a combination of both performance related data and cash flow level data was used, as well as fund-level statistics. The Appendix provides additional detail on the employed data supplementation process.

4.2.1 Independent Variables

The set of predictors was selected based on the reviewed literature and the available data. For macroeconomic drivers the GDP and Treasury bond were sourced from FRED (Federal Reserve Economic Data), while the MSCI world index return was sourced from WRDS (Wharton Research Data Services). Table 2 provides an overview of the selected variables.

Table 2: Overview of the independent variables

The table provides the name and a brief description of the selected predictors used in the analysis, as well as the type of each variable. The predictors for which the squared term was also included in the analysis, have a note included in their description.

Variable	Description	Type
fund_size	The size of commitments to a PE fund (squared term also included).	Numerical (Continuous)
fund_no_overall	The number of funds raised by the fund manager (squared term also included).	Numerical (Discrete)
fund_no_series	The number of funds raised by the fund manager in a specific series (squared term also included).	Numerical (Discrete)
geo_diversified	Indicates whether the fund invests in firms located in a single or multiple countries.	Binary
ind_diversified	Indicates whether the fund invests in firms operating in a single or multiple industries.	Binary
VC_specialization	Indicates whether the fund invests in firms in a specific financing stage or firms in different financing stages (VC funds only).	Binary
geo_focus	Indicates whether the fund invests primarily in NA, EU or Other.	Categorical (3 categories)
GDP_yoy	The nominal (YoY) growth rate if the US GDP in the vintage year of the fund.	Numerical (Continuous)
DGS10	The yield of a 10-year US Treasury bond in the vintage year of the fund.	Numerical (Continuous)
MSCI_World_yoy	The annual return of the MSCI world index in the vintage year of the fund.	Numerical (Continuous)
funds_raised_in_VY	The number of funds raised in the vintage year of the fund.	Numerical (Discrete)

4.2.2 Dependent Variable

In this thesis PME was selected as the appropriate metric for fund performance. However, the availability of data allows for a similar analysis to be performed for the IRR and TVPI metrics, as well. Out of all the available PME implementations, the original KS-PME was chosen, based on the reasons discussed in the literature review section. For some of the funds in the sample Preqin provided the KS-PME values. For others the values were calculated from the cash flow level data (detailed procedure described in the Appendix). The formula used for KS-PME calculations is as follows:

The Future Value (FV) at a given date *n* is calculated for all distributions and contributions (Cash Flows) of a fund:

$$Future\ Value\ \ = (Cash\ Flow)_t \times \frac{(S\&P500\ Value)_n}{(S\&P500\ Value)_t} \qquad (3)$$

$$for\ all\ t \in (0, n)$$

Where *n* is either the date when the fund is officially dissolved (for liquidated funds) or the date of the last reported Net Asset Values (NAV) (for closed funds). KS-PME is subsequently calculated as:

$$KS - PME = \frac{\Sigma\ \mathrm{FV(Distributions)} + NAV_n}{\Sigma\ \mathrm{FV(Contributions)}} \qquad (4)$$

Where *NAV* is equal to zero for liquidated funds and reported by the GP for closed funds.

Since the NAVs are reported by the GP, questions about the validity of the reported values arise. The paper by Brown et. al. (2019) investigates if PE funds manipulate reported returns. They conclude that underperforming managers inflate reported returns, but are less likely to raise subsequent funds, while top performing funds understate valuations. The index value used in the KS-PME calculation is the 'close' value of the S&P500 index, sourced from the WSJ. If the transaction occurred outside of trading days, the last available index value was used.

The final dataset consists of 1434 funds, out of which 721 employ a BO strategy and 713 employ a VC strategy. Their vintages range from 1985 to 2017. Funds raised after 2017 were excluded due to not completing the majority of their investments and as such the validity of their performance forecast would be questionable. Moreover, 72% of the funds invest primarily in NA, followed by 16%, which invest in EU and the remaining 12% invest in other regions. Furthermore, 435 of the funds in the sample are liquidated, while 999 are closed. The collective committed capital for the funds in the sample amounts to 1.58 trillion USD. Table 3 contains the descriptive statistics of the sample.

Table 3: Descriptive statistics

The table provides the descriptive statistics of the VC and BO samples, as well as the entire sample pre-split. The value in the parenthesis below each category is the standard deviation. All the statistics are provided for the sample post-preprocessing i.e., the data that is directly used by the ML algorithms.

	VC	BO	All
Fund Size ($Mn) mean	448.72	1749.86	1102.92
Fund Size ($Mn) median	280.0	752.5	425.5
	(651.68)	(2753.66)	(2108.17)
Fund No. Overall mean	4.74	5.0	4.87
Fund No. Overall median	4.0	4.0	4.0
	(3.98)	(5.29)	(4.68)
Fund No Series mean	3.91	3.78	3.84
Fund No. Series median	3.0	3.0	3.0
	(2.80)	(2.26)	(2.55)
PME mean	1.16	1.16	1.16
PME median	0.97	1.11	1.05
	(1.05)	(0.49)	(0.82)
TVPI mean	1.92	1.76	1.83
TVPI median	1.5	1.63	1.59
	(2.07)	(0.81)	(1.57)
IRR (%) mean	11.74	13.34	12.54
IRR (%) median	8.6	13.2	11.4
	(35.35)	(18.10)	(28.04)
Called (%) mean	95.48	94.10	94.79
Vintage mean	2007	2009	2008
No of Funds	713	721	1434

4.3 Data Preprocessing

Data preprocessing includes transforming or otherwise preparing the data so it can be interpreted and parsed by learning algorithms. Prior to data preprocessing the sample contained 6211 funds, out of which 5471 had calculated IRRs, 5833 had calculated TVPIs and 698 had calculated PMEs. After the data supplementation procedure, the sample consisted of 1475 funds, all containing their respective performance values. Subsequently, funds with missing relevant predictor data were excluded, resulting in a sample of 1434 observations.

4.3.1 Constructing Categorical Variables

Preqin provides the location and industry in which the funds' investments are focused. Moreover, it provides the information about the financing stage of the investment companies of the fund, as well as which strategy the fund employs. The large number of categories in the geographical focus variable provided by Preqin makes it impractical to use as is. Consequently, I reduced the number of the funds' geographical focus categories to EU, NA and Other, which is in line with the analyses conducted in past literature. Furthermore, I categorized the industry diversification variable as *diversified* if the fund invested in companies located in multiple industries and *non-diversified* if it invests in companies located in a single industry. Similar categorization was employed for the geographical diversification variable and the VC specialization variable (used for VC funds only). For categorizing the fund strategy variable, I relied on prior research, which classifies the strategies as either BO or VC (see Appendix for detailed explanation of the categorization process), and accordingly separated the dataset.

4.3.2 Encoding Categorical Variables

Encoding categorical data i.e., converting the data into numerical values, is necessary to ensure the proper functioning of the ML algorithms. Moreover, the performance of many ML algorithms is dependent on the encoding procedure used for categorical variables. There exist many different ways in which categorical data can be encoded, and the technique used is dependent on what type of categorical data the sample contains. Categorical data can be classified as ordinal or nominal. The data is ordinal, if the categories can be ordered in some way (e.g., low, medium, high), while the data is nominal, if such ordering is not possible (e.g., Norway, USA, Brazil). The sample of funds, used in the analysis contains nominal data, with up to three categories. Consequently, I selected binary encoding if a given variable

consisted of two categories, and one-hot encoding if a given variable consisted of three categories. To encode the dependent variable, I constructed a custom encoder, which sets the *KS-PME* variable as 1 if its value exceeds a predefined hurdle rate, and as 0, otherwise. The hurdle rate for this analysis was set to 1, to separate the funds which have outperformed the public equity market, from those who did not. However, the code allows for an arbitrary hurdle rate. Table 4 contains categorical variables, their respective categories, and their encoding.

Table 4: Categorical variables

The table provides the result of applying the encoding methods and the type of encoders used for each of the categorical variables.

Variable	Categories	Encoding
geo_diversified	1 – diversified 0 – non-diversified	Binary
geo_focus	[1, 0, 0] – EU focused [0, 1, 0] – NA focused [0, 0, 1] – Other	One-Hot
ind_diversified	1 – diversified 0 – non-diversified	Binary
VC_specialization	1 – specialized 0 – non-specialized	Binary
KS-PME	1 – exceeds hurdle rate 0 – subceeds hurdle rate	Binary

If the variable has two categories, binary encoding is equivalent to label encoding i.e., encoding each category as a positive integer (e.g., 0 – Cat, 1 – Dog, 2 – Cow, 3 - Chicken). However, if the variable has more than one category, one-hot encoding is used for nominal data and label encoding is used for ordinal data. The reason for this selection is that label encoding nominal data might artificially introduce a relationship between the different categories (i.e., 1 is less than 2 and 2 is less than 3), and as such might cause the ML algorithm to malfunction.

4.3.3 Transformations and Feature Scaling

Before the ML algorithms are applied, the sample's numerical data is usually transformed. The reason for the transformation is to either help the data become more interpretable, for the data to meet the assumptions of inferential statistics, to

ensure the proper functioning of a ML algorithm, or to deal with potential outliers, which worsen the out-of-sample performance of the model. The selection of models used on the sample data includes linear models, which perform better if the data is normally distributed. Consequently, I performed log-transformations on the *fund_size* and *fund_no_overall* right-skewed variables to make them approximately conform to normality. Moreover, the *fund_no_overall* variable contained outliers, whose effects I reduced by Winsorizing the data to the 99-percentile range. To ensure the proper functioning of the distance-based ML algorithms (e.g., kNN), I standardized the data using the z-score normalization method. Additionally, the standardization process improves the efficiency of most ML algorithms, by increasing the speed of learning and thus leading to faster convergence. The z-score normalization method is implemented by applying the following formula to each observation of the sample's numerical data:

$$z = \frac{x - \bar{x}}{\sigma} \tag{5}$$

Where z is the new value, x is the old value, $\bar{x}$ is the mean of the data, and σ is the standard deviation of the data.

The result of applying the method transforms the data so the mean of the values is zero and the standard deviation is one. Importantly, the order in which the data transformation methods are performed is paramount. Firstly, either log-transformation or Winsorization should be performed. Since the log-transformation is monotonic, Winsorizing the data before or after leads to the same result. Secondly, standardization (or other data rescaling techniques) should be applied. The reason behind this ordering is that performing log-transformation after standardizing the data would either not be possible (log-transforming negative values is impossible) or would defeat the purpose of the standardization procedure (log-transforming the data would result in it no longer having a mean of zero and standard deviation of one). The standardization procedure should also strictly be done after splitting the sample into the test and training split. The reason is that if standardization is performed before, the test split might contain information about the training split i.e., the mean and standard deviation of the entire sample.

Consequently, the estimation of the out-of-sample performance, would not really be out-of-sample.

5 Research Methodology

To demonstrate the feasibility of predicting PE returns, I used several principally different ML models: Logistic Regression, k-Nearest Neighbours, Linear Support Vector Classifier, Support Vector Machine, Decision Tree, and Random Forest. Moreover, I used CV to 'tune' the hyperparameters of each of the aforementioned models, as well as to improve the estimate of the models' out-of-sample performance.

5.1 Cross Validation Implementation

After the sample data has been appropriately prepared for use in the ML models, it must first be divided into the Test and Training set. The optimal size of each set is not clearly defined, however, for small datasets, as the one used in this thesis, the recommended split is 15% Test and 85% Training. Consequently, I split my original datasets (BO and VC) according to this recommendation. The reason for the Train/Test division is that we want to accurately evaluate the ML models' performance i.e., ensure that the models are general enough to be applied to data, which the algorithm has not 'seen' before. Consequently, we use the Training set to, intuitively, train the model and the Test set to evaluate its performance strictly after all the training is completed. Additionally, a Validation set is usually extracted from the Training set. It is used to 'tune' the models' hyperparameters and evaluate their performance. This technique of evaluating a ML model by training it on the subsets of input data (Training set) and evaluating it on the complementary subsets of data (Test and Validation set) is CV. There are many ways in which CV can be implemented (e.g., Hold-out Method, Shuffle Split Method, Leave-One-Out Method etc.). The appropriate method is selected based on the type of input data (e.g., time-series data requires a different CV approach than cross-sectional data) and the task requirements (e.g., medical ML applications require a more complex/exhaustive CV approach than marketing ML applications). For the purposes of this thesis, I selected the Stratified K-Folds Method with six folds.

5.1.1 Stratified K-Folds Method

The K-Folds CV Method is implemented by randomly splitting the Training set into K unique datasets, with (generally) the same number of samples. The model is subsequently trained on K-1 datasets and evaluated on the Kth dataset. The process

is repeated until all the folds are used in the evaluation and training. The result is a selection of scores, which are calculated based on the model's performance in each of the evaluation folds. These scores can successively be used in the hyperparameter selection process, or they can be averaged to get an estimate of the model's out-of-sample performance. The scoring method used in this thesis is accuracy scoring, whose loss function is the zero-one loss function:

$$L(i,j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i,j \in M \tag{6}$$

Where M is the set of class labels (In the case of this thesis over or underperform). The function returns 0 as many times as the model classifies the objects correctly ($i = j$) and 1 as many times as the model classifies the objects incorrectly ($i \neq j$). Therefore, the accuracy score of a model for which the loss function returned '1' seven times and '0' three times would be 30%.

The stratified K-folds CV Method replaces the random sampling of data into folds with stratified sampling. This ensures that each fold includes an approximately equal ratio of labels. In the case of this thesis, that would mean that each fold includes an approximately same number of funds that have outperformed the public equity market. The reason for selecting this specific method is that it is not as computationally intensive, while still being an improvement over the traditional Leave-One-Out Method. Furthermore, since the purpose of the evaluated ML models is not deployment, but instead more of a proof of concept, complicated CV approaches are unnecessary. Moreover, the stratified method was used instead of the traditional one, because the dependent variable's distribution is skewed. Consequently, using the stratified method over the traditional one leads to better model performance.

5.2 Model Overview

In the following model operation descriptions n will be used to indicate the number of observations in the sample. In the case of this thesis, n = 1434 i.e., the total number of funds. Furthermore, p will denote the number of independent variables. In the case of this thesis p = 13 for the BO dataset, and p = 14 for the VC dataset.

Moreover, the values of the independent variables will be represented in a matrix X:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \quad (7)$$

Where the value of the independent variable $j \in [1, p]$ for fund $i \in [1, n]$ is x_{ij}.

The values of the dependent variable will be represented as a vector y:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad (8)$$

Where the value of the dependent variable for fund $i \in [1, n]$ is y_i. In the case of this thesis $y_i \in \{0, 1\}$.

5.2.1 Logistic Regression

The Logistic Regression model is the simplest linear parametric model used to solve binary classification problems. It is an adjustment of the Linear Regression model, in a way that limits the possible prediction results between zero and one. It does this by changing the fitting function into the sigmoid (also called logistic) function:

$$h(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (9)$$

Where $X = (X_1, \ldots, X_p)$ are the predictors, $p(x)$ is the predicted probability that the dependent variable is equal to one, and $\beta = (\beta_0, \ldots, \beta_p)$ are the coefficients we are trying to estimate.

Due to the model's structure, it can be fitted (i.e., its coefficients estimated) using LS. However, the more general method of maximum likelihood is preferred, due to its favourable statistical properties. The intuition behind the ML method in the case of this thesis is that it tries to find the coefficients, which result in a probability that is close to one (zero) for all funds that outperformed (underperformed) the public

equity market. This is implemented by choosing the coefficients, which maximize the likelihood function:

$$l(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i:y_i = 1} p(x_i) \prod_{i':y_{i'} = 0} (1 - p(x_{i'})) \quad (10)$$

Importantly, the classification of fund, based on the predicted probability can be selected appropriately. The default setting is that the funds whose probability of outperforming the public equity market is greater or equal than 0.5, are classified as outperforming and others as underperforming.

The advantages of using a simple linear model are that its results have high interpretability (e.g., a one-unit increase in *fund_size* is associated with an increase in the log odds of *PME > 1* by β_{fund_size} units). Moreover, the model's simplicity enables it to perform better on data with few observations and/or data whose signal-to-noise ratio is low (which is the case in this thesis and often the case with financial data). However, overfitting might occur, as a consequence of including too many predictors in the model. I mitigate this problem in two ways. Firstly, I only include features selected based on the reviewed literature. Consequently, I have a strong reason to believe that all included independent variables provide important predictive value to the model. Secondly, I use a regularization technique to penalize the inclusion of predictors which do not adequately improve the fit. Regularization methods reduce the variance (at the cost of an unequal increase in the bias) of the coefficient estimates by shrinking them towards zero. Consequently, by choosing the appropriate amount of penalization, we can achieve the minimum estimated out-of-sample error. Because of the belief of the importance of all included predictors, I selected the Ridge Regularization method (also called L2 Norm Regularization). This method allows for shrinkage of the coefficients towards zero, but not setting them to zero (i.e., not excluding them). The amount of penalization is regulated via the parameter λ. If λ is set to 0, then there is no penalization and if λ is set to infinity, then all the coefficient estimates will approach zero. To select the amount of penalization that leads to the lowest estimated out-of-sample error, I used 6-Fold Stratified CV. The disadvantage of using this model is that it is not flexible. Consequently, if the underlying assumption of linearity does not hold, the model

will perform poorly. Furthermore, capturing interaction effects between the predictors is possible only if they are included in the independent dataset, which becomes unviable with an increasing number of predictors.

5.2.2 K-Nearest Neighbours

K-Nearest Neighbours is the simplest non-parametric model, which can be used to solve binary classification problems. It classifies a given observation x_0, by estimating the conditional probability that the observation belongs to a given class, based on the response values of the K datapoints 'closest' to it. In the case of this thesis that would mean a fund is predicted to outperform the public equity market, because the funds whose predictors exhibited similar values, also outperformed the public equity market. Mathematically, the conditional probability is estimated as:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in A_0} I(y_i = j) \quad (11)$$

Where j is a class (In the case of this thesis *outperforming*), A_0 is the set of K points 'closest' to observation x_0, and $I(y_i = j)$ is the zero-one loss function. Intuitively, the observation x_0 is subsequently assigned to the class with the largest probability (The boundary probability for classification is 0.5 by default and was not changed in the analysis).

The advantage of using this model is that the method of finding the appropriate solution is very intuitive and therefore simple to explain. Moreover, it is computationally fast and the model itself is very flexible. Therefore, it can be used to estimate a decision boundary of any form. The main drawbacks of using this method is its low interpretability and the curse of dimensionality. With an increasing number of predictors, it becomes increasingly more difficult to find points, which are close to each other. In the case of this thesis that would mean that a fund cannot be classified, since no other fund in the dataset exhibited predictor values similar to those of that fund. The bias-variance tradeoff of kNN can be regulated via the K parameter. If K is equal to one, the variance is the highest (the bias is the lowest) and there is severe overfitting. If K is equal to n then the variance is the lowest (bias is the highest) and the classification is simply determined by the number of observations in the sample that belong to a given class. Consequently,

the value of K is selected so that the model achieves the highest possible out of sample performance. I used 6-Fold Stratified CV to estimate the out-of-sample performance of models with different values for K and selected the one with the highest accuracy score.

5.2.3 Support Vector Classifier and Support Vector Machine

The SVC (also known as a Linear Kernel SVM) is a linear parametric ML model, which classifies the observations in the sample dataset via the use of a hyperplane. A hyperplane is a $p - 1$ dimensional space that is used to separate the observations. Therefore, if an observation lies on one side of the hyperplane it is placed in one class and vice versa. The hyperplane is positioned so that the perpendicular distance from the closest observations to the hyperplane is maximal. Consequently, only the closest observations to the hyperplane determine its position and orientation and are thus appropriately renamed to Support Vectors. The mathematical representation of the SVC is:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, x_{i'}) \quad ; \quad K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{12}$$

Where β_0 and α_i are the coefficients, we are trying to estimate, S is the set of the support vector points and, $K(x_i, x_{i'})$ is the Linear kernel function i.e., the inner products of the observations. The mathematical explanation as to why only the inner products of support vector observations affect the linear classifier $f(x)$, is beyond the scope of this thesis.

Since all the observations are not always linearly separable, the SVC allows for some to be positioned on the wrong side of the hyperplane. The number and severity of the violations is determined by the tuning parameter C. If C is equal to zero then no violations are allowed, which, if the separation is possible, leads to the highest possible fit to the Training data. However, this solution is not very robust since a single observation sufficiently close to the existing hyperplane can severely impact its position and orientation. Thus, the tuning parameter C controls the bias-variance tradeoff. If C is small, then variance is high (bias is small) and accordingly if C is large the variance is small (bias is high). The advantages and disadvantages of this

ML model are similar to those of the Logistic Regression and the prevention of overfitting is handled in an equivalent way as well.

SVM extend the SVC by allowing for non-linear boundaries and can thus be implemented to solve non-linear classification problems. It achieves this by enlarging the feature space i.e., including more features. While the feature space can be enlarged simply by adding squares, cubes etc. to the Linear Kernel function, the computationally viable alternative is to use different kernel functions. Consequently, the $K(x_i, x_{i'})$ function in equation (12) can be changed based on the sample data and user preferences (e.g., Polynomial Kernel, Sigmoid Kernel etc.). For the purposes of this thesis, I selected the Radial Kernel, whose function is:

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right) \quad (13)$$

Where γ is a positive constant, which regulates the gradient of the decent of the Radial Kernel function and can be 'tuned' via CV to achieve the lowest estimated out-of-sample error. The advantages of SVM over SVC is that the model allows for a greater degree of flexibility and can achieve a better performance when dealing with non-linear classification problems.

5.2.4 Decision Tree and Random Forest

Decision Trees are non-parametric models that classify the observations in the data sample by stratifying or segmenting the predictor space into a number of simple regions, based on a set of splitting rules. There exist several different splitting rules, based on which the algorithm automatically decides on the predictor and the value of that predictor, upon which it splits the set. The most basic rule is the classification error rate minimization. It entails minimizing the number of the Training observations in a specific region that do not belong to the most common class. In the case of this thesis, that would mean creating splits that group together the funds that have outperformed or underperformed the public equity market, based on the values of the predictors. However, the classification error rate minimization is often not sufficiently sensitive. Consequently, other measures are preferable (e.g., Cross-entropy, Log loss). For the purposes of this thesis, I selected the Gini index criterion,

which is a measure of total variance across the two classes i.e., node purity. The tree grows by repeating the splitting process on each of the subsequently created regions. The depth of the tree i.e., the total number of splits made, is controlled by the hyperparameter *d*. Intuitively, the depth hyperparameter also controls the bias-variance tradeoff. If *d* is small the variance is low (bias is high) and vice versa. Therefore, the depth of the tree is selected via CV. The advantage of using decision trees is that they can capture both linear and non-linear predictor relationships. Nonetheless, linear models tend to perform better in case of linear relationships. Furthermore, they are very simple to explain and closely mirror the human decision-making process. However, in practice they often do not display the same level of predictive accuracy as other classification approaches. Moreover, they have inherently high variance i.e., they often fail to generalize, since a small change in the training data can result in a very different set of splits. The main reason for this is the propagation of the change in a split to all the splits below it.

RFCs significantly improve the performance of DTCs by producing multiple trees, which are combined to yield a single consensus prediction. In the classification setting the prediction is made by obtaining a class vote from each tree, and then classifying using a majority vote. When building each tree of the forest, the number of features, the features themselves, and the data used is randomly selected from the Training set. Consequently, the correlation between the trees is reduced and the model is less sensitive to the original dataset. The number of trees generated by the algorithm and the maximum depth of each tree is controlled hyperparameters. In practice the number of trees should be as high as will still improve the model and the depth should be enough to achieve the desired number of observations for each node split. Using the RFC offers many advantages. They have very high interpretability, and they often perform remarkably well with very little parameter tuning required. Furthermore, they can capture complex interaction structures in the data and can be used to estimate linear and non-linear decision boundaries.

5.3 Model Comparison

5.3.1 Confusion Matrix

After the models have been trained on the Training data, their performance is assessed using the Test data. The performance assessment is based on the number of correct and incorrect predictions generated by the model.

These results are generally represented in a Confusion Matrix:

$$\begin{bmatrix} & Positive\ (PP) & Negative\ (PN) \\ Positive\ (P) & \text{True Positive } (\mathbf{TP}) & \text{False Negative } (\mathbf{FN}) \\ Negative\ (N) & \text{False Positive } (\mathbf{FP}) & \text{True Negative } (\mathbf{TN}) \end{bmatrix} \quad (14)$$

Where *P* and *N* are the number of observations in the Test data, with positive and negative outcomes (funds which have outperformed or underperformed the public equity market), respectively. Moreover, *PP* and *PN* are the number of positive and negative predictions of the outcomes, based on the observations in the Test data. *TP* (*FN*) is the number of positive outcomes that were predicted correctly (incorrectly). *TN* (*FP*) is the number of negative outcomes that were predicted correctly (incorrectly).

From the results of the confusion matrix, different metrics of model performance can be derived:

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

$$Precision = \frac{TP}{TP + FP} \quad (16)$$

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (17)$$

For the purposes of this thesis, precision is the most important metric. This is because we want to ensure that the funds, we invest in, will overperform the desired benchmark i.e., we want to minimize investment mistakes. Consequently, sensitivity is not as important since it essentially measures the 'missed opportunities' for fund investment. The relatively higher importance of precision

compared to recall is a desired characteristic of our ML use case, since there is usually a tradeoff between a high precision and a high sensitivity in practice.

5.3.2 ROC Curve

A different way of evaluating binary classification model performance is by measuring the area under the Receiver Operating Characteristic (ROC) Curve. The ROC Curve is constructed by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). TPR is simply another term for Sensitivity, while FPR can be derived from the Confusion Matrix as follows:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \quad (18)$$

The ROC curve can be interpreted as a representation of the costs (FP) and benefits (TP) of using a particular model. The perfect classification model would have a FPR equal to zero and TPR equal to 100%, while a random guess i.e., a model that has a 50% chance of correct classification (e.g., a coin flip), would be represented as a diagonal line from the origin to the 100% TPR and FPR points. Consequently, the points above the diagonal represent good classification results and points below represent bad results. To rank the effectiveness or predictive power of a model more simply, the area under the ROC Curve is used. The larger the Area Under the Curve AUC the better the model. While the appropriateness of the AUC as a performance measure has been questioned (e.g., by Hanczar et. al. (2010)) it is still extensively used in the ML community.

6 Analysis and Results

6.1 Model Selection

Given the different investment strategies, styles, and operations of VC and BO funds all the models were applied to datasets comprised of the funds of each type. Furthermore, the probability threshold used to classify the funds was 50%. Consequently, the funds who were expected to outperform the benchmark with a 50% probability or higher were classified as *outperforming* and those with a lower probability were classified as *underperforming*. However, the models allow for the adjustment of the threshold according to user preferences. To ease the model selection process, I constructed a scoring system which assigns a final score to the model by equally weighing the Accuracy, Precision, and AUC scores. The performance of the models is summarised in Table 5:

Table 5: ML model performance

The table provides the performance results based on the appropriate metrics discussed in Chapter 5.3. All the results were calculated using the Test dataset. CV represents the mean of the CV folds accuracy scores, ACC represents accuracy, as given by equation (15); PCS represents precision, as given by equation (16); and AUC is the performance measure described in Chapter 5.3.2. The abbreviations for model names are given in the List of Abbreviations. FS is the final score of the model, constructed by equally weighing the aforementioned scores.

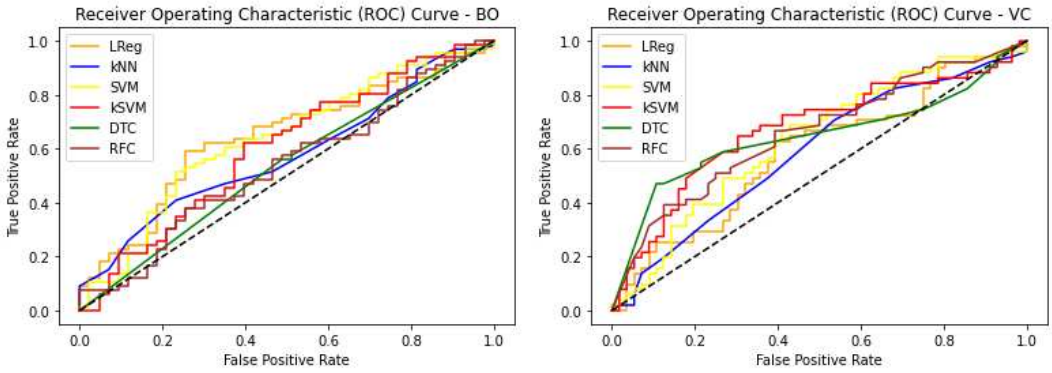
Model	BO					VC				
	CV	ACC	PCS	AUC	***FS***	CV	ACC	PCS	AUC	***FS***
LR	*0.62*	0.60	0.64	0.64	***0.627***	*0.56*	0.59	0.57	0.59	***0.583***
kNN	*0.60*	0.61	0.63	0.58	***0.607***	*0.58*	0.56	0.54	0.59	***0.553***
SVC	*0.64*	0.63	0.63	0.64	***0.633***	*0.54*	0.53	0.55	0.63	***0.570***
SVM	*0.61*	0.63	0.64	0.60	***0.623***	*0.58*	0.65	0.65	0.67	***0.657***
DTC	*0.60*	0.53	0.63	0.53	***0.563***	*0.59*	0.66	0.67	0.64	***0.657***
RFC	*0.59*	0.53	0.60	0.54	***0.557***	*0.57*	0.63	0.63	0.66	***0.640***

For the BO dataset, SVC is the best performing model, while for the VC dataset DTC and SVM are the top performers. Overall, all of the models, in both datasets have a greater classification capability than the random classifier. Furthermore, the BO dataset model's scores are on average similar than that of the VC dataset, indicating that the selected predictors contain information about BO funds and VC

funds. In the BO dataset, linear models are the best performers, indicating that the relationship between the selected predictors and fund performance is linear or, more likely, that the data's Signal-to-Noise Ratio is low and therefore simpler models perform better. In the VC dataset, non-linear models are the best performers, indicating that the relationship is non-linear, or that they are able to capture complex effects, which linear models are unable to. Surprisingly, the performance of RFC is lower than DTC in the VC dataset. Given that RFC improves upon the DTC approach the expected results is the opposite. Figure 1 shows the ROC curves of the models of both the datasets:

Figure 1: ML model ROC curves

The figure illustrates the ROC Curves for the BO and VC dataset. The dotted diagonal line represents the ROC of the random binary classifier.

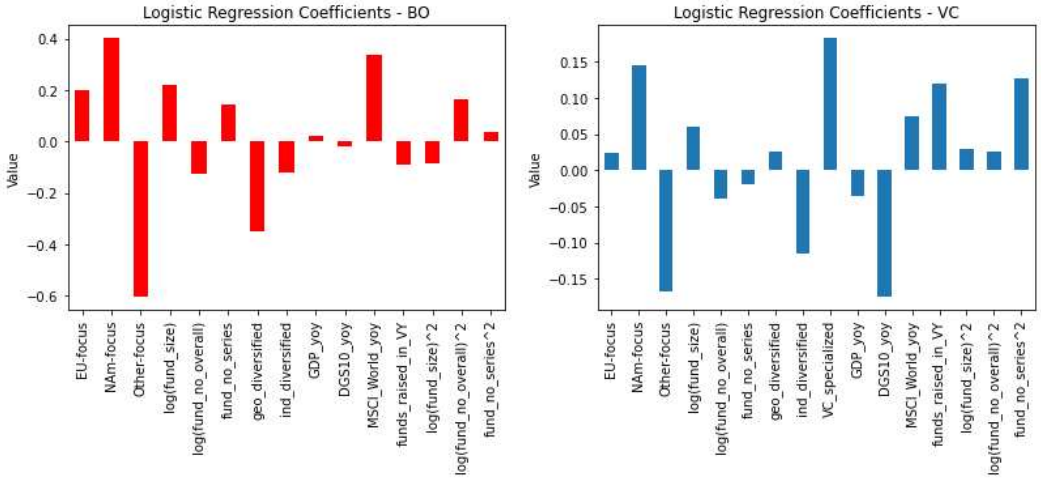


6.2 Predictor Analysis

The primary aim of this thesis was to test the viability of using ML to assist in PE fund investment decision making process. However, the interpretability of results is an important and desired model characteristic and will consequently be discussed as well. From the selected model set, LR, SVC and RFC offer insight into the importance of the selected predictors. Since LR and SVC have similar solution approaches, the SVC coefficient results will not be discussed. Because LR is a parametric model, the insight into predictor importance can be gained through model coefficient estimates. The LR model coefficient values are illustrated in Figure 2:

Figure 2: Logistic Regression coefficient estimates

The figure illustrates the parameter values for the BO and VC dataset. The x-axis contains parameter names while the y-axis contains their values. The bar plots were generated using the Training data i.e., the same subset of data that was used to train the model and obtain results in Table 5.



The coefficients for geographic focus show similar results for both BO and VC funds. However, the effect of European fund focus is proportionally lower in VC funds than BO funds, which might be attributed to a lower percentage of European funds included in the VC dataset (see Appendix). Furthermore, the effect of industry diversification for both VC and BO funds is shown to be negative, which is inconsistent with the findings of Lossen (2006), who finds a positive relationship to fund performance, and with Aigner et. al. (2008) who does not observe any relationship. Moreover, geographical diversification is shown to have a negative effect for both VC and BO funds, for which Lossen (2006) and Aigner et. al. (2008) find no effect. The VC Specialization (i.e., diversification across financing stages) coefficient's positive effect is consistent with the findings of Lossen (2006). The effect of fund size for BO funds is consistent with the findings of Kaplan and Schoar (2005), Roggi et. al. (2019), and Harris et. al. (2022), which identify a concave relationship to performance. However, for VC funds the relationship suggested by the LR coefficients is convex. Furthermore, manager experience, measured by *fund_no_overall* and *fund_no_series* demonstrates differentiating effects for BO and VC funds. Firstly, *fund_no_overall* has a negative effect on fund performance for both VC and BO funds. Moreover, it exhibits a concave relationship. The negative effect is conflicting with the results with the work from Kaplan and Schoar (2005), Roggi et. al. (2019), and Harris et. al. (2022). Secondly, *fund_no_series* is

shown to have a positive effect on performance for BO funds and a negative effect for VC funds. The positive effect is in line with prior research, while the negative effect is not. The effects of macroeconomic performance drivers for BO funds correspond to the findings of the reviewed literature. Additionally, the macroeconomic drivers, except for *GDP_yoy*, also exhibit effects in line with the literature.

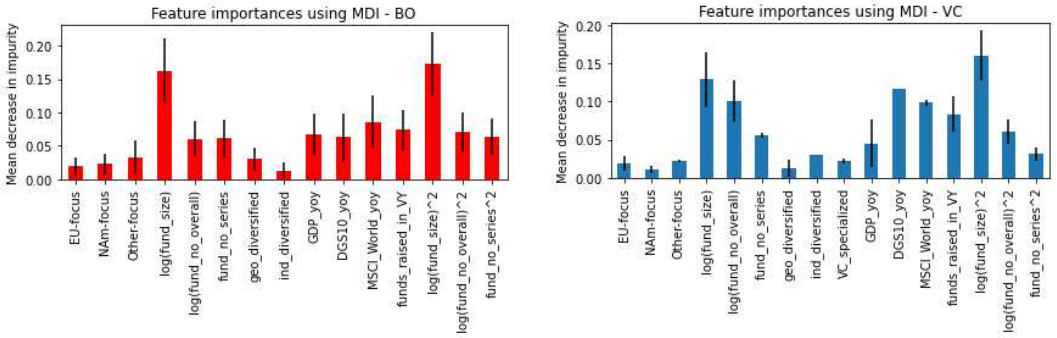
The features with the highest (relative) predictive power for both datasets are *Other_*focus, *VC_specialized*, and *DGS10_yoy*.

There are several different explanations as to why funds, which invest outside of NA or EU are predicted to be less successful. Firstly, they consist of newer funds, which are consequently not liquidated and therefore underestimate their residual value (Brown et. al., 2019). Secondly, PE markets in these areas are less developed, and as such administrative, travel, and other associated costs might be higher. Finally, regulatory constraints in these countries might inherently cause the returns to be lower. As mentioned in the Literature Review section, the general hypothesis of diversification is that it harms returns. The reason for this is the high degree of information asymmetry and agency problems present in the industry. Consequently, having specialized knowledge of the companies in a particular financing stage is assumed to be beneficial in the PE firms' portfolio company selection process. The positive effect of this predictor implies that these costs outweigh the benefits of diversification. The reason interest rates have a negative impact on performance might be that in times when interest rates are high, the number of companies which are able to use debt financing decreases. As such they seek alternative ways of financing i.e., rely on PE firms. Consequently, the number of 'good' investment opportunities increases, and the associated performance of PE firms is higher. The degree of predictive power *DGS10_yoy* has on VC funds, compared to BO funds is surprising, since debt financing is usually associated with BO fund performance. The reason for this might be the higher percentage of NA funds in the VC dataset, for which the chosen interest rate proxy is more appropriate.

The importance of predictors as per the RFC is given in Figure 3:

Figure 3: Random Forest feature importance

The figure illustrates the importance of features in the RFC for the BO and VC dataset. The x-axis contains parameter names while the y-axis contains the mean decrease in impurity. The bar plots represent the feature importances of the forest, along with their inter-trees variability, which is represented by the black error bars. The bar plots were generated using the Training data i.e., the same subset of data that was used to train the model and obtain results in Table 5.



The RFC calculates feature importance based on how much that feature is used in each tree of the forest. The importance of a feature is computed as the meaned reduction of the criterion, brought by that feature. In the case of this thesis the Gini Criterion is used to evaluate the splits, which can also be interpreted as a measure of node purity. Thus, the vertical axis of the plots in Figure 3, represent the mean decrease in impurity. In the RFC the higher the column in Figure 3, the more important the feature is deemed to be e.g., the most important feature for both datasets is the fund's size and its square.

The economic reasons behind the *fund_size* feature's importance are several. A fund's size determines how many portfolio companies a PE firm can have. Subsequently, a firm with more portfolio companies enjoys the benefits of diversification. However, the firm with fewer companies can pay more attention to its individual investments. Moreover, larger funds have less difficulty dealing with the fixed costs associated with running a PE fund, compared to smaller funds.

7 Conclusion

The intention of this thesis was to examine the viability of using ML tools to assist in PE investors investment decision making process. The chosen metric for fund performance assessment is the KS-PME, due to its ability to reflect risk adjusted returns and its rising popularity among LPs. Consequently, the fund selection question was translated into a binary classification problem. The funds were classified based on whether they outperformed a user specified PME benchmark. Those that outperformed the benchmark were classified as *outperforming* and those that did not were classified as *underperforming*. The analysis was conducted using the KS-PME = 1 benchmark, which translates the problem into a selection of funds which are predicted to outperform or underperform the S&P500 public equity index. The ML models chosen to take on this binary classification problem, were trained and tested on data ranging from fund-level statistics to macroeconomic data. Moreover, all the data used in was acquired from either commercial or publicly available data sources. Relevant predictors were selected based on prior research on PE performance drivers as well as data availability. Due to the difference in the investment strategies of VC and BO funds, the original dataset was appropriately split. Consequently, all the models were trained on separate datasets. The analysis showed promising results for both VC and BO funds, with the top performing models in each dataset reaching 63% and 66% accuracy, respectively. For the BO dataset, linear parametric models performed better than non-parametric models, which can most likely be attributed to the privation of data. For the VC dataset, non-linear models performed better, suggesting the presence of complex effects, which simpler models are unable to capture. The findings indicate a possibility of using ML as a complementary tool in the PE investment decision making process. They allow investors to scour many investment opportunities and perform detailed due diligence on only the most promising ones. Furthermore, they can adjust the desired return (benchmark) based on their own risk profile.

While the research conducted in the thesis was intended to be proof of concept, there are several improvements that can be made to enhance the ML model performance. Firstly, several datasets sourced from different data providers can be used to minimize the sample selection bias. Moreover, proprietary data and

predictors can be included to further personalize the fund selection process. Secondly, the values of the used predictors were based on the vintage year of the fund (e.g., funds raised in the vintage year). Consequently, some of the used information may not be available to investors at the time of fund raising and the predictions would be incorrect. Therefore, an improved ML model would be trained on the data available on investors when the investment is made. Finally, due to the long holding periods of PE funds, a large proportion of funds included in the sample is not liquidated. Consequently, the NAVs, reported by the GPs, used in the PME calculation might be erroneous. As such, using a dataset which contains fewer non-liquidated funds can offer a significant improvement.

References

Gu, S., Kelly, T. B., Dacheng, X. (2019). Empirical Asset Pricing via Machine Learning. *Chicago Booth Research Paper No. 18-04; 31st Australasian Finance and Banking Conference 2018; Yale ICF Working Paper No. 2018-09*.

Gompers, A. P., Gornall, W., Kaplan, N. S., Strebulaev, A. I. (2016). How Do Venture Capitalists Make Decisions?. *Stanford University Graduate School of Business Research Paper No. 16-33*, *European Corporate Governance Institute (ECGI) - Finance Working Paper No. 477/2016*.

Phallipou, L., Gottschalg, O. (2009). The Performance of Private Equity Funds. *The Review of Financial Studies,* 22(4), 1747-1776

Long, M. A., Nickels, J. C. (1996). A Private Investment Benchmark. *AIMR Conference on Venture Capital Investing*, 1-17

Kaplan, N. S., Schoar, A. (2005). Private equity performance. Returns, persistence, and capital flows. *Journal of Finance*, 60(4), 1791-1823

Korteweg, G. A., Nagel, S. (2013). Risk-Adjusting the Returns to Venture Capital. *NBER Working Paper No. 219347*.

eVestment. (2017). Enhancing Private Equity Manager Selection with Deeper Data.

Sorensen, M., Jagannathan, R. (2013). The Public Market Equivalent and Private Equity Performance. *Columbia Business School Research Paper No. 13-34*.

Harris, R. S., Jenkinson, T., Kaplan, N. S., (2014b). Private Equity Performance: What Do We Know?. *The Journal of Finance*, 69(5), 1851-1882.

Fenn, W. G., Liang, N., Prowse, S. (2001). The Private Equity Market: An Overview. *Financial Markets, Institutions & Instruments*, 6(4), 1-106

Phallipou, L., Zollo, M. (2005), What Drives Private Equity Fund Peformance?. *Working Papers – Financial Institutions Center at The Wharton School*, 1-29

Lossen, U. (2006). The Performance of Private Equity Funds: Does Diversification Matter?. *Munich Business Research Working Paper Series No. 2006-14*.

Kaserer, C., Diller, C. (2009). What Drives Private Equity Returns? – Fund Inflows, Skilled GPs, and/or Risk?. *European Financial Management*, 15(3), 643-675

Aigner, P., Albrecht, S., Beyschlag, G., Friederich, T., Kalepky, M., Zagst, R. (2008). What drives PE? Analyses of Success Factors for Private Equity Funds. *The Journal of Private Equity*, 11(4), 63-85

Roggi, O., Giannozzi, A., Beglioni, T. (2019). Private equity characteristics and performance: An analysis of North American venture capital and buyout funds. *Economic Notes*, 48(2)

Harris, S. R., Jenkinson, T., Kaplan, N. S., Stucke, R., (2022). Has Persistence Persisted in Private Equity? Evidence from Buyout and Venture Capital Funds. *Fama-Miller Working Paper*.

Hutchinson, M. J., Lo, W. A., Poggio, T. (1994). A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks. *The Journal of Finance*, 49(3), 851-889.

Lo, W. A., Mamaysky, H., Wang, J. (2002). Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *The Journal of Finance*, 55(4), 1705-1765.

Gavrishchaka, V. V., Banerjee, S. (2006). Support Vector Machine as an Efficient Framework for Stock Market Volatility Forecasting. *Computational Management Science*, 3, 147-160.

De Spiegeleer, J., Madan, D. B., Reyners, S., Schoutens, W. (2018). Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. *Journal of Quantitative Finance*, 18(10), 1635-1643.

Farquad, M. A. H., Bose, I. (2012). Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, 53(1), 226-233.

Imandoust, S. B., Bolandraftar, M. (2013). Application of K-Nearest Neighbour (KNN) Approach for Predicting Economic Events: Theoretical Background. *Journal of Engineering Research and Application*, 3(5), 605-610

Butaru, F., Qingqing, C., Clark, B., Sanmay, D., Lo, W. A., Siddique, A. (2016). *Journal of Banking and Finance*, 72, 218-239.

Fischer, T., Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669.

Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., Chen, H. (2017). An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach. *Computational Economics*, 49, 325-341.

Cheng, D., Pasquale, C. (2018). A Reinforced Urn Process Modelling of Recovery Rates and Recovery Times. *Journal of Banking and Finance*, 96, 1-17.

Freyberger, J., Neuhierl, A., Weber, M. (2018). Dissecting Characteristics Nonparametrically. *Fama-Miller Working Paper, Chicago Booth Research Paper Np. 17-32*.

Warin, T., Stojkov, A. (2021). Machine Learning in Finance: A Metadata-Based Systematic Review of the Literature. *Journal of Risk and Financial Management*, 14, 302.

Brown, W. G., Gredil, R. O., Kaplan, N. S., (2019). Do private equity funds manipulate reported returns?. *Journal of Financial Economics* 132(2), 267-297

Samuel, L. A., (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 44, 206-226.

Mitchell, M. T. (1997). Machine Learning. *McGraw-Hill.*

Bergadano, F., Gunetti, D. (1996). Inductive Logic Programming: From Machine Learning to Software Engineering. *MIT Press*.

Brown, W. G., Harris, S. R., Jenkinson, T., Kaplan, N. S., Robinson, T. D. (2015). What Do Different Commercial Data Sets Tell Us About Private Equity Performance?. *SSRN working paper 2701317*.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, 26, 822-830.

Bain & Company. (2022). Global Private Equity Report. https://www.bain.com/insights/topics/global-private-equity-report/

Fraser-Sampson, G. (2010). Private equity as an asset class. *John Wiley & Sons, Ltd.*

James, G., Written, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. *Springer*.

Hastie T., Tibshirani, R., Friedman J. (2017). The Elements of Statistical Learning Data Mining, Inference, and Prediction. *Springer*.

A APPENDIX

A.1 Data Supplementation Process

For the data supplementation process, a combination of two datasets was used. The first one is the Fund-level dataset, which contains predictor variable's data and performance data. The second one was the Cash-Flow-level dataset, which contained information about the distributions and contributions of funds as well as basic fund-level data information.

The reason the data supplementation procedure was necessary is because of the missing data in the Fund-level dataset. The details are provided in Table 6.

Table 6: Fund-level data - missing data breakdown

The table provides the information on the missing values of the Preqin provided fund level data. The data was not subjected to any processing or filtering, except for the limitation of the vintage to 2017 or younger and fund strategy to *Buyout*, *Turnaround*, *Venture Capital*, and *Growth*. The data is provided for the dataset after removing the duplicates.

	Present	Missing
Fund Size	5895	282
Fund No. Overall	6123	54
Fund No Series	6066	111
KS-PME	694	5483
TVPI	5801	376
IRR	5440	737
Vintage	6177	0
Strategy	6177	0
Status	6177	0
Geographic focus	6133	64
Industries	6141	36

Originally the Fund-level dataset contained 6211 funds. After removing the duplicates, the dataset contained 6177 funds.

The Cash-Flow-Level dataset contained the information of 2410 funds. However, the dataset did not contain the information on important predictors (e.g., Fund No Series), and could therefore not be used in the analysis. Consequently, the missing performance data in the Fund-Level dataset would have to be supplemented by the performance data calculated from the Cash-Flow-Level dataset.

To calculate the KS-PME value, equation (4) was used. For the IRR calculation I used the *pyxirr* library, which calculates the IRR in the same way as Excel and Preqin. However, since the IRR is not calculation is not always possible, the IRR values were missing for some funds. The TVPI was calculated using the procedure described in Section 3.1.2.

Since the Cash-Flow-Level dataset did not necessarily include all of the transactions (i.e., distributions and contributions) that a fund has made, the performance metrics would have to be 'close enough' for the supplementation to be possible. Therefore, if the PME/TVPI/IRR value from the Cash-Flow-Level dataset was within a certain threshold of the corresponding value from the Fund-Level dataset, the supplementation was allowed, otherwise the observation was dropped. The thresholds selected were absolute, and their values were 0.05 for PME and TVPI and 0.5% for IRR.

The details of the supplementation rules are provided in Table 7.

Table 7: Supplementation rules

The table provides the information the conditions which had to be satisfied for the supplementation of the performance metrics in the Fund-Level dataset, with those calculated using the values in the Cash-Flow-Level dataset. TVPI, PME, and IRR_{fl} represent the existence of these performance metrics in the Fund-Level dataset. IRRcf represents the existence of IRR in the Cash-Flow-Level dataset. If the number below the given threshold is 1, that signals that the value exists in that dataset. However, if it is 0, that signals it does not. The right column represents the action, written in bold, and the reasons for the action, where applicable.

TVPI	IRR_{fl}	PME	IRR_{cf}	
0	0	0	0	**Drop** – No Values Exist
0	0	0	1	**Drop** – No Values To Compare
0	0	1	0	**Drop** – Missing IRR
0	0	1	1	No Funds Satisfy Condition
0	1	0	0	**Drop** – Missing IRR
0	1	0	1	**Check IRR**
0	1	1	0	**Check PME**
0	**1**	1	1	**Check PME and IRR**
1	0	0	0	**Drop** – Missing IRR
1	0	0	1	**Check TVPI**
1	0	1	0	**Drop** – Missing IRR
1	0	1	1	**Check TVPI and PME**
1	1	0	0	**Check TVPI**
1	1	0	1	**Check TVPI and IRR**
1	1	1	0	All Values are Available
1	**1**	1	1	All Values are Available

After the performance metric supplementation procedure, all the funds with missing predictor data were dropped, resulting in a final sample size of 1434 funds.

A.2 Data Categorization Process

The categorical predictors used in the analysis contained an unpractical number of categories when sourced directly from Preqin. Consequently, I grouped some of the categories to achieve the categorization structure described in Table 4. The geography related variables were constructed from the **Geographic Focus** variable from the Fund-Level dataset. For the *geo_diversified* variable, the grouping of categories is described in Table 8.

Table 8: Geographic diversification category grouping

The table provides the description of the grouping used to construct the categories of the *geo_diversified* variable. If the category was grouped into the (Non-)Diversified group it is written under the (non-)diversified column.

Diversified	Non-Diversified
North America	US, France, Israel
West Europe	South Africa, Germany
Asia	Canada, Thailand, Mexico
Europe	UK, Italy, China, Turkey
South Asia	Australia, Japan, India
Nordic	Brazil, Poland, Lithuania
Central and East Europe	Greater China, Indonesia
North Africa	Finland, Peru, Portugal
Americas	New Zealand, Denmark
Sub-Saharan Africa	Russia, South Korea
Africa	Netherlands, Spain
East and Southeast Asia	Switzerland
South America	
Australasia	

For the *geo_focus* variable, the grouping is described in Table 9.

Table 9: Geographic focus category grouping

The table provides the description of the grouping used to construct the categories of the *geo_focus* variable. If the category was grouped into the NA, Europe it is written in the respective column. If the category is not in either NA or Europe, it was classified as Other.

NA	Europe
North America	West Europe, Europe
US	Nordic
Canada	Central and East Europe
	France, UK, Italy
	Poland, Lithuania
	Finland, Portugal
	Denmark, Russia
	Netherlands, Spain
	Switzerland
	Germany

For the *ind_diversified* variable, the grouping was done similarly to the *geo_diversified* variable.

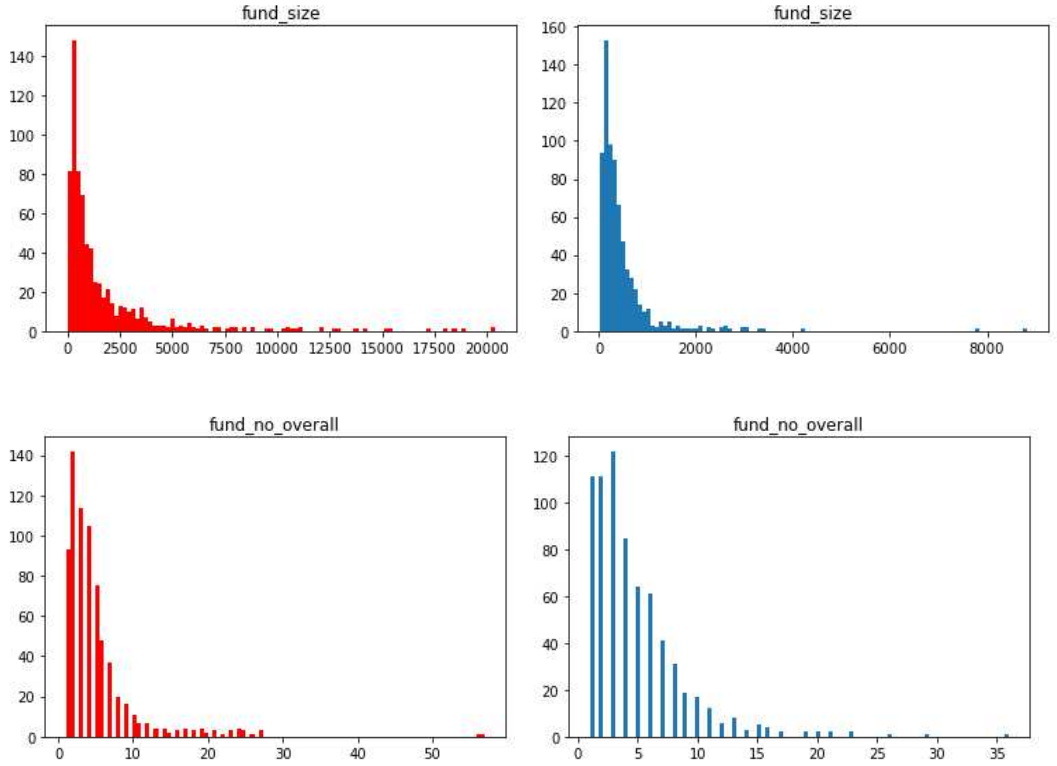
In the literature, a funds strategy is classified as either VC or BO. However, Preqin recognizes additional strategies, which were subsequently grouped in the aforementioned categories. The grouping approach was the same as the one used by Brown et. al. (2015). The VC group encompassed the following strategies used by Preqin: *Growth*, *Venture (General)*, *Early Stage: Seed*, *Early Stage*, *Early Stage: Start-up*, *Expansion / Late Stage*. Consequently, the VC funds were categorized as non-specialized, if their strategy was *Venture (General)* and specialized if their strategy was one of the others available. The BO group encompassed the following strategies used by Preqin: *Buyout*, *Turnaround*.

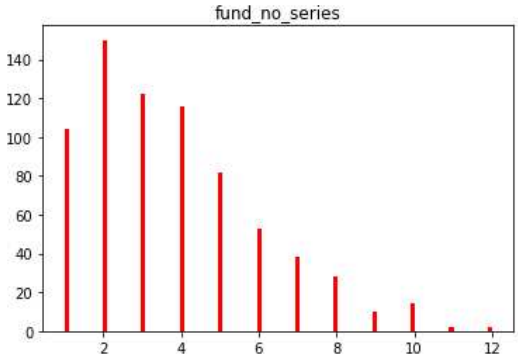
A.3 Distributions and Statistics of the Predictive Variables

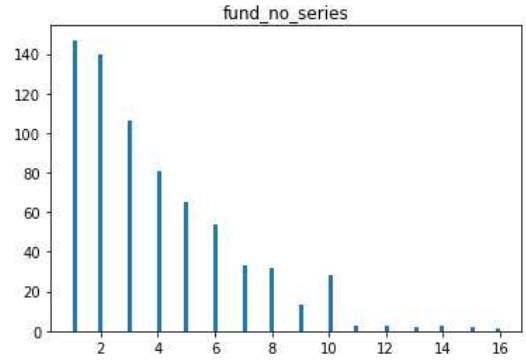
The distributions of the predictor variables are given below. The histograms of the BO dataset are on the left and are given in red, while the histograms for the VC dataset are on the right and are given in blue.

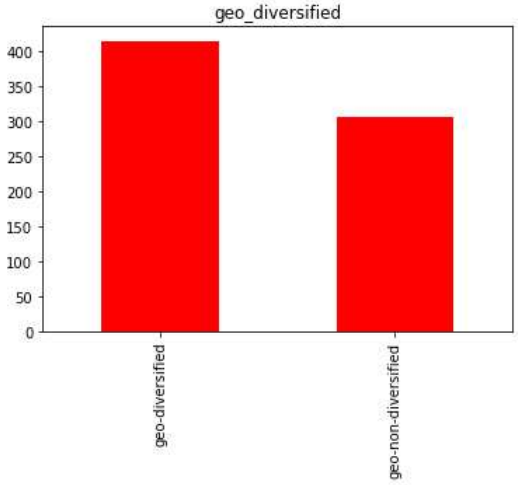
Figure 4: Predictive variable distributions

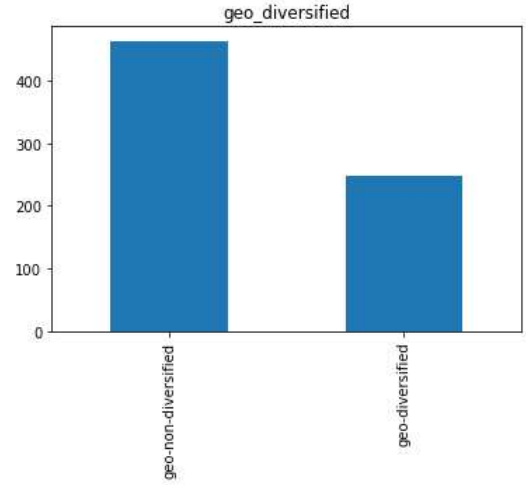
The figure illustrates the histograms of the predictive variables in the analysis. The histograms are given before any transformations were applied on the features.

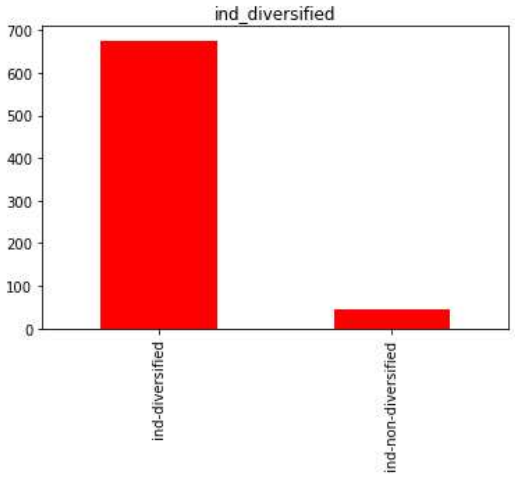


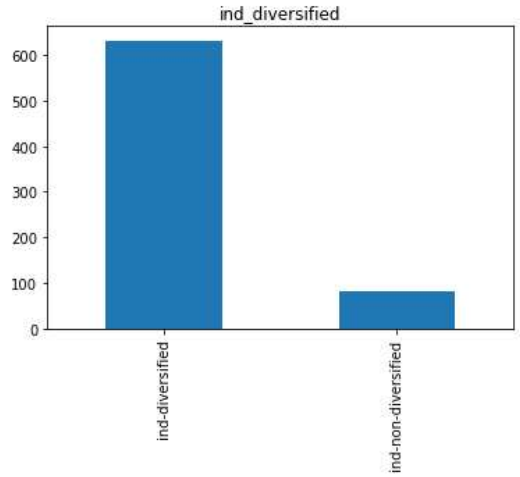

fund_no_series
0
20
40
60
80
100
120
140
2
4
6
8
10
12

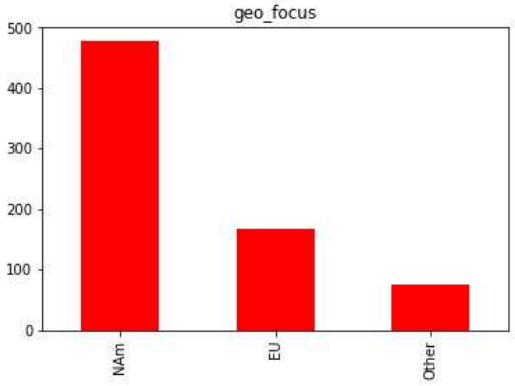

fund_no_series
0
20
40
60
80
100
120
140
2
4
6
8
10
12
14
16

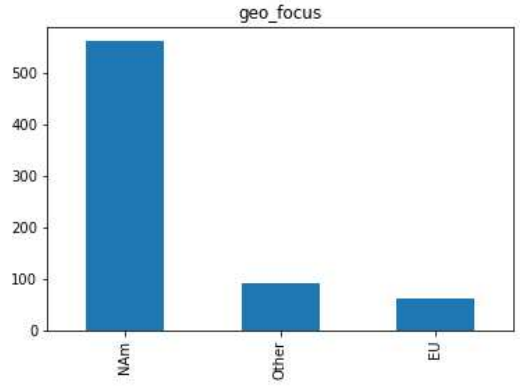

geo_diversified
0
50
100
150
200
250
300
350
400
geo-diversified
geo-non-diversified


geo_diversified
0
100
200
300
400
geo-non-diversified
geo-diversified


ind_diversified
0
100
200
300
400
500
600
700
ind-diversified
ind-non-diversified


ind_diversified
0
100
200
300
400
500
600
ind-diversified
ind-non-diversified


geo_focus
0
100
200
300
400
500
NAm
EU
Other


geo_focus
0
100
200
300
400
500
NAm
Other
EU

The macroeconomic predictor statistics for BO and VC funds are given in Table 10 and Table 11, respectively

Table 10: BO fund macroeconomic variable statistics

The table provides the descriptive statistics of the macroeconomic variables used in the analysis of BO funds.

	min	max	mean	median	stdev
GDP_yoy	-1.98	7.85	4.17	4.19	1.76
DGS10	-1.84	1.22	-0.17	-0.16	0.53
MSCI_World_yoy	-22.17	30.88	9.87	14.35	13.83
funds_raised_in_VY	1	106	75.31	78	26.49

Table 11: VC fund macroeconomic variable statistics

The table provides the descriptive statistics of the macroeconomic variables used in the analysis of VC funds.

	min	max	mean	median	stdev
GDP_yoy	-1.98	7.85	4.45	4.20	1.72
DGS10	-2.95	1.22	-0.14	-0.13	0.54
MSCI_World_yoy	-22.17	30.88	10.07	15.51	15.12
funds_raised_in_VY	1	106	70.19	78	27.88

A.4 Hyperparameter Analysis

The hyperparameters for the models were selected by using 6-Fold Stratified CV for each parameter value. Consequently, the parameter who achieved the highest accuracy score was selected. The hyperparameter plots for the models where a single hyperparameter was tuned are given below. The tuning for the (VC) BO dataset is presented on the (left) right in (blue) red.

Figure 5: Hyperparameter tuning

The figure illustrates the plots of the accuracy scores for different given hyperparameters values. The plots are displayed for the models for which a single hyperparameter was tuned.