Private Equity Fund Selection
- A Machine Learning Approach

Simran Pachnanda * Rishi Raj **

Stockholm School of Economics

May 17, 2021

Abstract

This paper aims to build a prediction model using machine learning (ML) algorithms to offer decision support for private equity investors - limited partners (LPs) in their fund selection process. Past literature has studied a range of factors that appear to drive the performance of private equity funds; some of these factors are known to LPs during fundraising such as targeted fund size, management experience, fund specialization level, state of the industry, and the overall economy. We tap into the predictive power of these factors by using them to train a range of supervised machine learning models in a binary classification setting that predicts the probability of a fund exceeding a predetermined performance threshold. We use the Public Market Equivalent measure developed by Kaplan and Schoar (2005) to construct our target variable which takes the value **1**, if the fund has a PME greater than one, and takes the value **0**, otherwise. Our models are based on a sample of 1058 Buyout (BO) funds and 659 Venture Capital (VC) funds sourced from Preqin. Our analysis shows some degree of performance predictability in both VC and BO funds with the top models reaching an accuracy of 69% for BO funds and 61% for VC funds. We also test the predictions of two of the models, Logistic Regression and Linear Discriminant Analysis (LDA) against a naïve investment strategy which also showed favorable results.

Supervisor: Tobias Sichert, Assistant Professor, Stockholm School of Economics

Examiner:Jungsuk Han, Associate Professor, Stockholm School of Economics

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

*41708@student.hhs.se, MSc Finance

**41664@student.hhs.se, MSc Finance

Acknowledgements

We would like thank our supervisor Tobias Sichert for his continual feedback, discussions, and support during our project. Special thanks to Christopher Rosenqvist and Stockholm School of Economics Institute for Research (SIR) for their support with our topic's ideation and the continuous encouragement and support during our research. We would also like to take this opportunity to thank all our faculties for their interesting courses and discussions that led us to establish our starting point and context. Lastly, we would like to thank our respective families and friends for their supportive thoughts during our thesis.

Contents

Abbreviations

ML Machine Learning

PE Private Equity

LP Limited Partner

GP General Partner

PME Public Market Equivalent

BO Buyout

VC Venture Capital

IRR Internal Rate of Return

NAV Net Asset Value

LR Logistic Regression

LDA Linear Discriminant Analysis

QDA Quadratic Discriminant Analysis

KNN K-Nearest Neighbour

SVC Support Vector Classifier

MLP Multi Layer Perceptron

AUC Area Under the Curve

ROC Receiver Operating Characteristics

Chapter 1

Introduction

Since the turn of the century, the Private Equity (PE) market has seen an exponential boom in funding, primarily from institutional investors like pension and endowment funds who are increasing re-allocating their portfolios towards alternative assets. This significant rise in PE investment can be attributed to the consistent out-performance of the industry as a whole relative to the public market as evidenced in recent studies by Robinson & Sensoy (2013a), Harris et al. (2015) [4], and Higson & Stucke (2012)[13]. However, it is important to note that the difference in performance between top and bottom quartile funds in each vintage year has been significant (See figure 1.1) – averaging around 13.15% between 2000 to 2016. With the global PE market reaching an all-time-high of $6.5 trillion assets under management (AUM) in 2019 - an increase of 170% since 2010 - and the number of active PE firms more than doubling (McKinsey 2020) [22], this top-bottom quartile performance gap has widened further. This considerably increases the cost of a bad selection for limited partners (LPs) and puts pressure on their reliance on the traditional methods of investment selection that are time consuming and highly dependent on human judgement. This opens doors for exploring whether new tools like artificial intelligence (AI), which is making significant inroads in various avenues of the financial industry, could prove to be a complementary tool for LPs in their fund selection process as well.

While PE firms have historically been slow to incorporate new digital tools into the decision-making process and relied primarily on investor relations for deal making, the last few years have seen a wave of change as the industry reaches maturity and the competition among fund manager stiffens. According to Bain's private equity report 2021 [2], many PE firms have already gone digital by employing AI, big data, and web-based analytics for making smarter and faster decisions about their portfolio companies

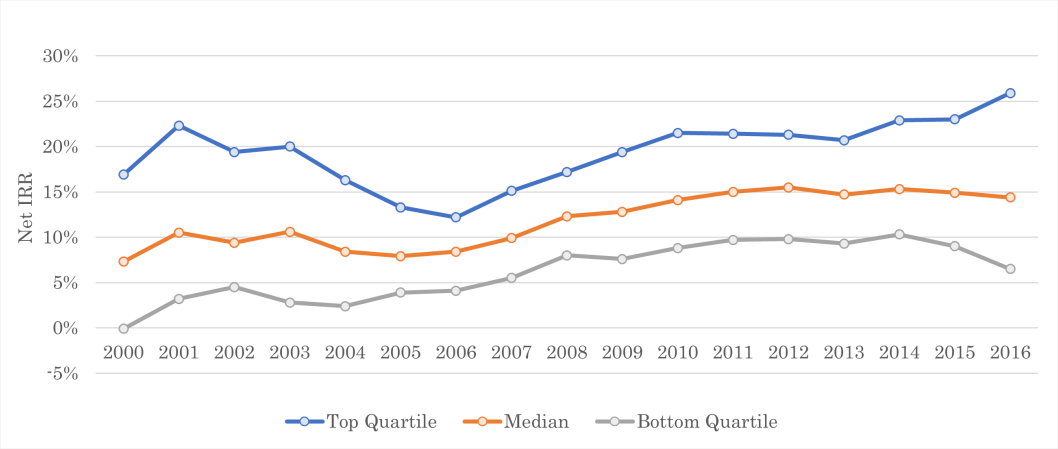


Figure 1.1: Private equity Net IRRs by Vintage Year (Source: Preqin)

and prospects. The need for incorporating digital aid into firm's due diligence process is becoming essential for the industry players to stay on top. One such example of leveraging AI is seen in the Swedish venture capital firm EQT Ventures's new AI-driven software called **Motherbrain** which the firm claims to have used for selecting more than 30% of their investment deals [8]. In this paper, we focus on exploring AI's applicability in the private equity industry from an investor's standpoint – that is the limited partners.

While historically many LPs simply invested in fund managers of previous top- quartile funds owing to the conventional wisdom of performance persistence in PE funds, a recent survey by eVestment (2017) [7] showed that only 19% of buyout funds raised after 2001 that were a successor to a top quartile performer repeated their performance. A similar study conducted by McKinsey (2017) [6] also supported this claim, observing a steady decline in top-quartile persistency in more recent vintages. This begs the question - what factors apart from past-fund performance are driving the returns on these funds and are any of them known to investors during the selection process?

While the ultimate performance of a fund is undoubtedly influenced by the decisions and conditions the manager faces during various stages of the fund's life cycle; empirical evidence suggests that at least some of these performance drivers are known to potential investors (LPs) at the time of fundraising. These include *fund-specific characteristics* like: the targeted fund size, management experience, financial and geographical scope, and *macroeconomics factors* like: interest rates, business cycles, level of competition and the overall state of the economy. We aim to tap into the predictive power of these variables

by taking a cross-sectional approach to predicting fund performance. These predictors have been chosen based on past literature on PE fund performance drivers and their availability. Our model is set up to follow a binary classification problem which aims to predict the likelihood of a fund beating a predetermined performance threshold. We use the Public Market Equivalent (PME) measure developed by Kaplan-Schoar [15] for constructing our target variable which takes the value **1**, if the fund has a PME greater than one and **0**, otherwise. A majority of our data, including the fund level cash-flows used for calculating our PME values, is sourced from Preqin's private equity database accessed via Wharton Research Data Services (WRDS). Given the nature of our problem, we focus on models known to work well with classification problems such as Logistic Regression with L1 and L2 penalty, Discriminant Analysis with Linear and Quadratic decision functions, Support Vector Machines, K-Nearest Neighbours and Neural Network. To the best of our knowledge, this is one of the first studies that applies machine learning into predicting private equity fund performance. We further address these three sub-questions during our study:

(1) Which models perform the best with our given data-set and why? Is the accuracy derived from advanced machine learning models significantly higher than the basic models?

(2) How accurate are these predictions overall and how do they fare when tested against naïve investment strategies?

(3) What features are driving the predictive power of our top performing models? Are they in line with the empirical evidence observed for them in past research?

Our analysis presents promising results for machine learning as a complementary tool in LPs' due diligence process. The top models for buyout funds – Logistic Regression, Linear Discriminant Analysis (LDA) and Support Vector Classifier (SVC) showed an overall accuracy of 69%, while the leading venture capital model K-Nearest Neighbour (KNN) showed a 61% accuracy. As for our predictors, most of them matched their

empirical findings, at least to the extent of the direction of their impact. However, some of them failed to exhibit a significant effect on fund performance. Some of the key drivers that stood out for BO included the level of PE activity in the year of fundraising measured by the number of PE funds raised that year, the prevailing yield on 10 Year Treasury Bonds and the targeted fund size. For VC funds, specialization in an industry presented to be a key driver of performance.

Thesis Outline

This thesis is broadly divided into seven chapters. In **Chapter 2** (Related Literature), we discuss the past literature on our topic, summarize their key findings and outline the gaps we are aiming to address. In **Chapter 3** (Theoretical Background), we provide the essential background readers require on private equity markets and machine learning. In **Chapter 4** (Data and Features), we give a description of our final sample of funds, followed by a detailed outline of how our target and predictor variables were sourced, cleaned, and compiled. In **Chapter 5** (Methodology), we provide the outline of our study's methodology that includes - description and characteristics of the various machine learning methods used in our analysis along with how we apply and cross verify these modelling techniques in practice. We also discuss evaluation metrics used for comparing the different models. In **Chapter 6** (Analysis and Results), we present the results and comparative statistics from our calibrated models. We then benchmark the top predicted funds by two of our ML models against a naive investment strategy. We conclude the chapter by discussing the key features driving our results. Finally, in **Chapter 7** (Conclusion), we summarize our main finding, discuss the limitation to our study and offer suggestions on how future research can improve in this area.

Chapter 2

Related Literature

This chapter discusses the past last literature related to our study. We have broadly divided the chapter into two subsections: The first talks about the literature related to private equity, in particular the empirical findings on what factors drive the returns of PE funds. The second, relates to the various areas of finance where machine learning has been applied and tested in the past. We conclude by pointing out the gaps in the interaction of these two literatures and how we aim to use the findings of our study to reduce this gap.

2.1 Private Equity

The conventional wisdom that "performance persistency" exists in the private equity (PE) industry has previously driven many LPs to top-quartile fund managers and shy away from new, untested funds. Early academic research that focused on buyout funds (BO) and venture capital (VC) raised in the 1980s and 1990s document strong evidence for this persistency (Kaplan and Schoar 2005 [15], Robinson et al 2016 [27]). Subsequent studies were conducted to investigate whether persistency weakened post – 2000 as the industry matured. Harris et al 2020 [11] looked into this question and found evidence to support the conventional wisdom for both pre and post 2000 funds. However, they also noted that since the capital raising period for a follow-up fund occurs about midway through the life of a GP's current fund, only an interim performance evaluation of the current fund is available to the investors which is based on the cash-flows occurred until that date and an estimated net asset value (NAV) of the unrealized investments. They found little or no evidence of performance persistency in BO funds when the information the LP actually has during fundraising– the interim and not the final performance of the

previous fund – was looked into. A rationale for this is offered by Jenkinson et al. (2013) [14] and Brown et al. (2019) [3] who study the interaction between interim valuations of current funds and subsequent fundraising. They show that while on average the fund's net asset values (NAVs) are conservatively reported, their valuations shoot up when the fundraising period of the follow-up fund (usually 3 to 5 years into the current fund) approaches.

A range of studies have been conducted that investigate factors that drive private equity apart from past fund performance. Some of these factors have inconsistent empirical findings between authors, further motivating us to explore their true relationship and predictive power. In our study, we focus only the firm-specific characteristics and macroeconomic variables that are available to LPs at the time of fundraising. Gottschalg et al.(2004) [25] and Kaplan and Schoar (2005)[15] were one of the first studies to explore features like fund size and management experience as potential drivers of fund performance. They observed a concave relationship for fund size and fund returns, implying the existence of an optimal fund size in terms of performance, beyond which the fund might start showing diseconomies of scale. They also conclude that more experienced GPs tend to raise better performing funds. Roggi et al. 2019[28] confirms these findings for fund size, however they suggest a convex relationship between experience and fund performance. Lossen (2006) [20] explored the impact of diversification in terms of financing stages, industries, and countries, on the fund's performance. His findings suggested that the return of a PE fund declines with diversification across financing stages and increases with diversification across industries. Aigner et al. (2008) [1] looked into the relationship between fund performance and firm-specific factors like: financing stage, experience of GPs, industry sector and certain exogenous factors such as: GDP growth, interest rate levels and the public markets environment. We provide a more detailed account of the empirical finding for each of our predictor variables in the **Data and Features** chapter.

2.2 Machine Learning in Finance

The exponential increase in availability of data and more affordable computing power has allowed technologies like artificial intelligence (AI) to flourish in the field of finance

research, particularly in the case its subsets machine learning (ML) who's models are known to perform well on noisy data-sets, which is a typical feature of financial data. One of the first references of ML in finance can be found in (Hawley et al 1990) [12] , who presents the applicability of neural networks as a tool for financial decision-making. Since then, ML has made significant inroads into the financial literature with research covering areas like bankruptcy prediction (Olmeda and Fernandez 1997 [23] , Zhao et. al. 2014 [30]), consumer credit risk modelling (Khandani, Kim Lo, 2010 [18]), understanding the default recovery rates (Cheng et al 2018) [5]; modelling investor sentiment (Renault 2017) [26] etc. In additional to these areas, research related to financial forecasting using ML has seen a boom in recent years, particularly in the case of cross-sectional stock market prediction (Kelly et al. 2019 [17], Gu et al. 2020 [10], Kozak et al. 2020 [19], Freyberger et al. 2020 [9]).

While many private equity players, particularly VC funds, are rapidly integrating their businesses with AI and machine learning, academic studies relating to this domain are still sparse. This is largely due to the difficulty in obtaining reliable and sufficiently large data-sets in PE in the past, given its reliance on voluntary reporting of data by LPs and GPs. However, commercial data-set providers like Pitchbook and Preqin which specialize in alternative asset classes like PE, are paving the way for machine learning driven research in these untested markets.

Chapter 3

Theoretical Background

3.1 Private Equity

Private Equity (PE) funds are investment houses that raises capital to invest in portfolios of non- publicly traded companies. This could be either a direct investment in a private company or buying out a public company and taking it private. Depending on their investment style, the equity stake they hold in these portfolio companies can range anywhere between 10% to 100%, with the aim to resell their share at a higher value on a future date. Two key investment styles in PE, which are also the two we will be focusing on in our study, are **Leveraged Buyout** (BO) and **Venture Capital** (VC). BO funds usually takes a controlling interest (50% to 100% equity) in mature companies and create value through active governance, operational improvements, and financial engineering. They typically finance their acquisitions though a significant portion of debt (60 to 90 percent of total capital) which gives them the name - leveraged buyout (Kaplan and Stromberg 2009) [16]. On the other hand, VC funds typically take minority stakes (around 20% to 50%) in start-ups and young companies that they believe have long-term growth potential.

3.1.1 Industry Structure

There are two key players involved in a private equity fund – general partners (GPs) and limited partners (LPs). The GPs are the managers of the fund – usually a PE firm - and are responsible for sourcing, acquiring, and managing the investments. The LPs are the fund's capital providers and are not actively involved in the day-to-day workings of the fund. They are usually either high net worth individuals or institutional investors

such as endowments, pension funds, sovereign wealth funds and insurance companies. The compensation structure for GPs has two main components – first a fixed annual management fee calculated as a percentage of committed capital (usually 2% but can range from 1.5 – 2.5 %). Second a variable interest - called the "carried interest" - that is paid as a percentage of the fund's profits (approx. 20 %) after the initial capital invested by the LPs is returned back. Additionally, there are some GPs that charge deal and monitoring fees to the companies in which they invest (Kaplan & Stromberg 2009) [16]. A more detailed analysis of the PE fee structure can be found in Robinson Sensoy (2012).

These funds are formed as partnerships or limited liability companies, with a fixed life span between 10 – 12 years. The initial four to five years are the "Investment Period" where the GPs invest the committed capital into companies, followed by the "Liquidation Period" where they exit their position from these companies and return the capital plus profits to the LPs. The exit strategies involve either a *secondary buyout* - reselling the company to another PE fund, a *strategic buyer* – selling to an industry competitor or an *Initial Public Offer (IPO).*

3.1.2 Measures of Performance

Traditional Methods

The Internal Rate of Return (IRR) has been the leading performance metric for comparing funds in the PE industry. Mathematically, it is defined as the discount rate that makes the net present value (NPV) of the investment equal to zero. The calculation is based on the fund's cash flows as well as the net asset value (NAV) at the time of the calculation if the fund is not liquidated. The draw-downs or "calls" are the negative cashflows while the distributions paid back to the LPs in the form of capital gains or dividends are the positive cashflows. It is a useful measure for comparing fund of the same vintage year since its calculation accounts for the issue of irregular timing and size of cash-flows present in this asset class. There are two different types of IRRs used in the industry: (1) The Gross IRR which is the return the fund makes on its investment before deducting fund expenses, management fee and carried interest. (2) The Net IRR that is the return the fund's LPs make net of all fees, interest, and expenses.

Despite it being the metric of choice in the industry, the reliability of this method has long been a point of discussion between industry specialists and academics alike due to some of its obvious pitfalls. Some of these include: (1) It assumes the distributions to the LPs are reinvested at the same rate of return as the fund's IRR at exit, which can over or understate the actual performance. (2) It does not take into account the scale of the investments. Thus, comparing investments with significantly different contributed capital can result in misleadingly high IRRs for smaller projects, even though the absolute gain (in terms of cash) is low. (3) It favours funds that have early exits even if their long-term performance does not match up to these early wins. A more detailed discussion about the pitfalls of IRR for the private equity industry can be found in Phalippou and Gottschalg (2009) [24].

One way of reducing IRR's pitfalls is by using other methods such as Money Multipliers[1] in parallel. These primarily include – DPI, RVPI and Net Multiple.

DPI (Distributions to Paid-in) is the proportion of the called-up capital that has been distributed or returned back to LPs.

$$DPI\ (\%) \ = \ \frac{Total\ LP\ Distribution}{Total\ LP\ Contribution} \ \times \ 100$$

RVPI (Residual Value to Paid-in) represents the amount at which an asset could be acquired or sold in a transaction between willing parties. This amount excludes any carry/performance fees earned by the GP and is shown as a percentage of total LP Contributions.

$$RVPI\ (\%) \ = \ \frac{Unrealised\ Value\ of\ Fund}{Total\ LP\ Contribution} \ \times \ 100$$

The Net Multiple is the ratio between the total value that the LP has derived from their investment – that is the distributed cash and securities plus the value of the LP's remaining interest in the partnership – and its total cash investment in the partnership.

[1] All definitions and formulas for money multipliers are taken directly from Preqin's Glossary to keep in line with the reported figures we use from their database.

This shows the scale of the return from an investment, which is not reflected in the IRR. Computationally, it is the sum of the DPI plus RVPI, expressed as a multiple:

$$Net\ Multiple\ =\ \frac{Distribution\ (\%)\ +\ Value\ (\%)}{Total\ LP\ Contributions}$$

It is important to note that while IRR and cash multiples collectively form a useful set of performance evaluators, they are still an absolute measure of performance. Given the operational differences in private equity market relative to other asset classes, their IRRs and multiples are not directly comparable. Moreover, they cannot be used for comparing funds raised in different vintage years as they do not control for factors that causes market wide movements.

Public Market Equivalent (PME)

Public Market Equivalents or PMEs are an alternative set of performance measures developed to benchmark the performance of a private equity fund against a public market index – for example in our analysis the S&P 500. The illiquid nature and irregular timing of cash flow make it difficult to compare private and public funds directly. Thus, the development of PME provides a more meaningful, "apples to apples" comparison for investors to compare different asset classes.

The fundamental idea in all PME methods is to calculate what the value of the fund's cash flows would be if they were invested in a stock market index instead. The capital calls are treated as money being invested into a stock index while the distributions to the LPs are taken as selling the stock market index shares. So in essence, it shows the market-adjusted multiple of invested capital (Harris et al. 2020)[11]. The first irritation of this measure was developed by Austin M. Long and Craig J. Nickels [21] in 1996 called the Long-Nickels PME. There have been many versions of it since then with minor adjustments aimed at overcoming the initial shortcomings. For our analysis, we use the version introduced by Kaplan and Schoar (2005) [15] called the Kaplan-Schoar PME or KS-PME. We explain the process of computing PME using this method in detail in the **Data and Features** section.

3.2 Machine Learning

The term Machine Learning (ML) was coined by Arthur Samuel in 1959 as a field of study that gives computers the ability to learn without being explicitly programmed. It is an approach to analyze data, and thereafter build and adapt models based on the same data such that the model is able to make a better prediction on unknown data. Despite ML being as a branch of computer science since late 60s and 70s, it has only recently found resurgence with advancement in technologies and recused computing cost.

Broadly, ML is classified into two categories: Supervised learning and Unsupervised learning, with additional subsets such as Semi-supervised and Reinforcement learning. We focus on the supervised machine learning models in our analysis.

Mathematically speaking, for an unknown relation or mapping function or the ground truth, we model it as

$$Y = f(x) + \epsilon \qquad ...(1)$$

where $'f'$ is unknown, and $'\epsilon'$ is the irreducible error. Now, we can never know the unknown population function $'f'$, but we can estimate it by fitting a model as below

$$\hat{Y} = \hat{f}(x) \qquad ...(2)$$

And, if we assume that $'f'$ is linear then we can write (2) as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p$$

where $\hat{\beta}_0, \hat{\beta}_1, ..\hat{\beta}_p$ are estimates of the for the true co-efficients.

The task of an analyst is to find the best estimates of the co-efficients by fitting several models and checking against some selected metric as per the business use case, for instance it can be R^2 in a regression setting or **F1-Score** in the classification setting. It must be noted that the accuracy of the estimates is related to the reducible error as it in our hand to find the best fitting $\hat{f}(x)$. We compute confidence interval in order to determine how close $\hat{Y}$ will be to the true $f(x)$ but because of $'\epsilon'$, the noise or the unexpected error, the

Prediction interval will always be greater than the confidence interval.

To summarize, ML is learning the function f that maps input variables X to output variable y. An algorithm learns this target mapping function from a training data. Different algorithms make different assumptions about the form of the function. Therefore, if we assume or simplify the function to a known form (closed form or fixed structure) then we call it as a Parametric method as discussed above. Linear models such as Linear regression, Logistic regression, and Support Vector Machines are typical examples of parametric methods. Similarly,if we don not make strong assumptions about the form of the mapping function then we land into the non-parametric world of methods. Non-parametric does not mean that they will not have parameters, but rather that the complexity of the model will grow with the increasing amount of training data so the number of parameters will grow proportionally. The typical examples are K-Nearest Neighbours (KNN) and Decision Trees.

3.2.1 Bias-Variance Trade Off

Bias of a statistical model is the estimation error between the actual value and the predicted value due to generalization whereas Variance of a statistical method refers to the amount by which $\hat{f}$ would change(variability of the model prediction) if we estimated it using a different training data set. For example, assuming a non-linear true model as linear will introduce high bias in the prediction for the unknown samples. A model is defined as flexible if it can fit into as many data points as possible. The trade off emanates from the flexibility of the model that is higher the flexibility higher the variance, and lower the Bias, that is a complex model. Similarly, simpler the model, higher the bias, and lower the variance. It is important to mention that such bias-variance trade off is controlled by hyper-parameters of the model such as λ in Logistic regression and C in Support Vector classifier, which are further elaborated in the Modeling section.Briefly, Under fitting can be understood as training error close to test error, i.e. High bias; Over fitting as low training error but with high variance.

Chapter 4

Data and Features

The quality of the data goes a long way in dictating how accurate a prediction can be. The reliability of the data provider, the frequency of data being reported, and the time-interval of the data were few of the questions we kept in mind before opting for our data source. We start this section by offering some descriptive statistics about our sample funds and summarize their historical performance. This is followed by a detailed description on how we sourced, cleaned, and compiled our final target and predictor variables along with the motivation behind each selection.

Our analysis primarily uses data sourced from **Preqin** – a commercial financial data and information provider on alternative asset markets. They collect majority of their data by putting in direct Freedom of Information Act (FOIA) requests to LPs and GPs for voluntarily making their information public which they complement with public filings and industry-recognized news sources. As of 2015, their data is sourced 38% from LPs, 59% from GPs and about 3% from public filings (Brown et al, 2015) [4].

4.1 Data Description

Our final dataset consists of 1717 funds – out of which 1058 are Buyout funds (BO) and 659 are Venture Capital (VC) funds. They range between vintage years[1] 1980 to 2016. We excluded any funds that were raised post vintage year 2016 to allow for only though funds who have completed the bulk of their investment period. Furthermore, we removed any BO fund with committed capital below $15 Million and any VC fund with committed capital below $10 Million, to only keep economically relevant funds in the dataset. A majority of funds in our sample are US based - around 85%, followed by mostly Euro-

[1]In our analysis, a vintage year is defined as the year in which the fund made its first investment

pean funds (10%). They collectively amount to a committed capital of 1.97 Trillion USD.

Table 4.1 presents some basic statistics of our sample. BO funds show an average fund size of $1149 Million relative to VC funds who raise relatively smaller funds averaging at $326 Million. The distribution of fund size is heavily right skewed for both BO VC funds. On average, about 95% of the committed capital was called for each fund, indicating that the fund size estimated at the time of fundraising is representative of the amount of capital ultimately invested in companies by the fund.

Table 4.1: Descriptive Statistics

	All Funds	Buyout (BO)	Venture Capital (VC)
Average Fund Size ($M)	1148.81	1661.1	326.3
Median Fund Size ($M)	425	725.2	232
	(2113.9)	(2548.8)	(345.4)
Average Net IRR (%)	12.7	13.9	10.8
	(23.7)	(13.1)	(34.4)
Average PME	1.18	1.20	1.15
Median PME	1.09	1.14	0.95
	(0.88)	(0.45)	(1.30)
Average Net Multiple	1.67	1.67	1.70
Average DPI (%)	121.7	121.4	122.3
Average RVPI (%)	45.24	43.97	47.27
Average Called %	94.7	93.6	96.4
No of Funds	1,717	1,058	659

*All values in parenthesis are standard errors

It is important to note that a majority of funds included in our analysis (over 80%) were raised during or after vintage 2000 which leaves our sample with a significant portion of closed funds (71%) relative to liquidated funds (29%) – see figure 3.1. This raises the concern about how reliable GP reported interim numbers are - in particular the Net Asset Values (NAVs) we use for calculating our target variable. While there is a possibility of NAVs manipulation by GPs to attract investors for successive funds, empirical evidence

suggests that these interim numbers are often conservatively reported relative to the final cash-flows and could be understated by as much as by 35% throughout the life of the fund (Jenkinson et al, 2013) [14]. The exception to this is the period when follow-on funds are being raised, usually around 3 to 5 years into the current fund's life, where valuations are often inflated to impress investors. These manipulations however, seldom go unnoticed by LPs and are taken as a negative signal (Brown 2019) [3]. Consequently, top-performing funds with an established tack record find little incentive in overstating their numbers, even during fund raising periods, given the risk of losing out on investors and tarnishing their reputation. Furthermore, the mark-to-market accounting standards - FAS 157 that came into force in 2006 requires PE funds to report thier balance sheet assets at a fair-value, thus further improving the accuracy of reported NAVs.

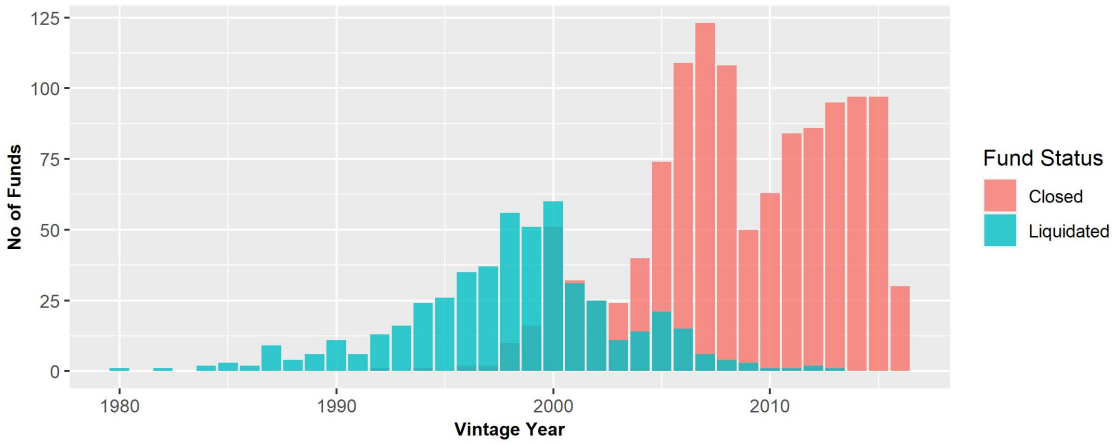


Figure 4.1: Distribution of Closed and Liquidated Funds

4.2 Data Selection

In this section we describe how we sourced and constructed our target and predictor variables, along with the motivation behind each selection.

4.2.1 Target Variable

Our model is calibrated as a binary classification problem, where the aim is to determine the probability of a fund exceeding a predetermined performance threshold. The target variable takes the value **1**, if the fund exceeds this threshold and **0** if they do not.

As discussed in the Theoretical Background section – while Internal Rate of Return (IRR) has been the performance measure of choice for the PE industry, we restrain from using it as our performance metric due to its well know limitations and its inability to capture changes in macroeconomics conditions – thereby not allowing for comparison between funds from with returns from different time periods. A solution for incorporating macroeconomics changes is by using IRR as a metric to rank funds of each vintage year into "quartiles" and then looking at the top quartile[2] funds in each vintage as our target. However, this still does not resolve the inherent limitation of IRR as a performance measure we discussed in earlier sections, thus we will be using a Public Market Equivalent (PME) method instead for constructing our target variable.

PME is steadily becoming a standard practice in the LP's due diligence process. According to a survey conducted by eVestment [7] in 2017 – 81% of respondents said they carry out PME analysis while 53% said they were expecting to increase their use of it. While there is no industry standard among the PME methodologies- with each having their own negatives and positives, the most popular choice among investors according to eVestment's survey was the Kaplan-Schoar PME (KS -PME), with 63% respondents using it. Sorensen and Jagannathan (2013) [29] provide a rigorous justification for the KS PME where they conclude that the measure holds valid regardless of the risk of PE investments provided the LPs have a log- utility preference[3]. We will be using KS PME as our method of choice for constructing the target variable.

Kaplan-Schoar PME

The KS PME is wealth measure that gives a ratio of gains to costs from the investment. A ratio of 1.2 would suggests that the fund, on average, outperformed the benchmark index by 20% – hence the LP benefited from investing in the fund relative to a similar investment they would have made in the public market. Conversely, if the ratio was 0.8, then the fund has under-performed on average relative to the market and the LP would have been better off investing in the stock index.

The calculation for the KS PME is as follows: At a given date **n**, a future value (FV)

[2]Top 25% funds
[3]That is investors are risk averse

is calculated for all the distributions and calls of the fund as follows:

$$Future\ Value\ \ (FV) = Cash\ Flow_t\ \times\ \frac{Value\ of\ the\ Stock\ Index\ at\ time\ n}{Value\ of\ the\ Stock\ Index\ at\ the\ time\ t}$$

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

For liquidated funds, **n** is the date when the fund is dissolved while for closed funds it is the date of the last reported NAV. The ratio is calculated as the sum of the FV of distributions plus the NAV at evaluation date **n** divided by the sum of the FV of calls.

$$PME\ =\ \ \frac{\sum FV(Distributions)\ +\ NAV_n}{\sum FV(Calls)}$$

For liquidated funds, the NAV is equal to 0, while for closed funds the NAV is calculated by taking the present value of the expected future cash flows.

Target Construction

For the fund level cash-flows, we use Preqin's Private Capital Cash Flow database - **(CASHFLOW)**. Our target variable is constructed to take the value **1**, if the fund had a PME greater than or equal to one, and takes the value **0** if the PME is less than one. Our motivation behind this threshold is that it would split the funds into groups that on average "beat the market" – thus making it a good investment for LPs - and those that do not. One key issue with any PME method is that the calculated values are sensitive to the benchmark index being chosen. To minimize the effect of this drawback, we calculate our PME values using two benchmark indices - **S&P 500** index as our primary index (obtained from CRSP via WRDS) and **Russell 3000** for robustness-check (obtained from Yahoo Finance). The binary nature of our target variable also minimizes this issue as long as the PME values from the two indices are not drastically different around the performance threshold of PME = 1. We observe that only 19 out of the 1717 funds show different values for their target variable and only 21 funds have an absolute difference in

PME greater than 0.1 between the SP 500 and Russell 3000. Hence, we believe that the PME values calculated using SP 500 are robust and we use those to construct our target variable.

4.2.2 Predictor Variables

Past research on performance drivers for PE funds have focused on liner and polynomial relationships. Our selected set of predictor variables builds on these empirical finding, many of whom show discordance of opinions among authors. We discuss these finding in four categories of variables below:

Fund Size: The past literature on the relationship between fund size and fund performance has been fairly divergent. Kaplan and Schoar (2005) [15] and Gottschalg et al. (2004) [24] observed a concave relationship between the variables, suggesting that beyond a certain level, an increase in fund size would not give an additional advantage to the fund in terms of performance and might even start affecting it negatively (diseconomies of scale). A subsequent study by Phalippou and Zollo (2005) [25] found evidence that support only a positive but not concave relationship between the two variable, while Lossen (2006) [20] and Aigner (2008) [1] suggest that the relationship might actually be negative. A more recent study by Roggi (2019) [28] however supports Kaplan and Gottschalg's findings of a concave relationship for both BO and VC funds. He suggests that this concave relationship could be interpreted as returns for small business being affected by their low bargaining power and high operating leverage while large funds suffer from dis-economies of scale, caused by the acceptance of less profitable investments, as well as agency and communication costs. Our variable: *Final_Size_USD*

Management Experience: Kaplan and Schoar (2005)[15] and Aigner (2008) [1] both observe that experienced GP tend to have a positive influence on the performance of a fund. However, according to the findings by Roggi (2019) [28] the positive effect of experience on performance only comes into play at very high sequence numbers for both BO VC funds – thus forming a convex relationship between experience and fund

performance. Two measure commonly used to proxy for management experience are: (1) **Firm's Age** - how long the GP has been in business or (2) **Fund Sequence** - the chronological number of the fund raised by the GP. We include both these measure in our analysis using variables - *Firm_Age*, *Fund_Number_Overall*, *Fund_Number_Series*

Specialization: There have been mixed evidence on whether specialization in a particular industry, financing stage (seed stage investing to mature companies) or country/geography positively affect the returns for PE funds. The supporters of the specialization hypothesis argue that a more focused PE firm would be in a better position to support its portfolio company as they understand the competition, technology, and market specific developments better. Lossen (2006) [20] explored the effect of all three of these specializations that suggested that while the return of a PE fund declines with diversification across financing stages, it increases with diversification across industries - indicating that the additional investment opportunities in new industries potentially outweighs the cost of industry diversification. He found no evidence for diversification across countries impacting the PE fund's returns. We believe the effects could be different now, given the significant growth in the PE industry across industries and geographies since this study was conducted. We use the following variables to explore these effects: *Geographic_Scope_Diversified*, *Industry_Diversified*, *VC_Specialized*

Macroeconomic Environment: The condition of the global economy and the PE industry at the time of fundraising plays a critical role in the overall performance of the fund. Aigner (2008) [1] interestingly observed a negative effect of Vintage year GDP and MSCI World Index growth on fund performance, suggesting that good economic conditions during the initial fund years potentially increase the prices for investment, thus lowering the overall return of the fund. He also found a negative effect of interest rate levels on return, which make intuitive sense especially for buyout funds who raised a sizeable portion of their investment capital using debt. We include all three of these parameters for the fund-raising year[4] in our analysis. For capturing the con-

[4]In our analysis, we are approximate the "Fund Raising Year" to be the year preceding the fund's vintage year. Hence, if the fund has a vintage year of 1990, we assume the fundraising took place in 1989

dition and prospects of the PE industry during fundraising, we are including the number of and percentage increase in new PE funds raised during the year prior to a PE fund's vintage. Our Variables: *GDP_yoy*, *TR_10yrs*, *yoy_MSCI*, *Funds_Raised_Last_Year*, *Pcent_Increase_Funds_Last_Year*

We additionally include variables that takes into account the location of fund raising (US vs Europe) and the geographic focus of the fund (US, Europe or Asia). While the relationship between the fund's performance and the factors listed above can differ depending on the fund's geographic focus and scope, our limited data set which is heavily skewed towards US based funds, and might be unable to capture these effects properly. However, the growing number of non-US based funds in the recent years urges us to still explore the effects of these variables. The complete list our predictor variables and their respective sources can be found in the Appendix A.2.

4.3 Data Processing

After having selected our target and predictor variables, the next step is to prepare the data for the modelling and analysis. This involves removing missing values, grouping and cleaning variables, and scaling the numerical features. We use the words - "features" and "predictors" interchanging in our analysis.

We started with a 3,278 PE funds reported in Preqin's *FUNDHISTORICPERFORMANCE* database which was reduced to a sample of 1717 funds after removing the following: (1) All funds with missing values for Net IRR or PME (2) All fund type that did not fit into the category of Buyout (BO) or Venture Capital (VC) funds - for example fund of funds, (3) Any fund with a vintage year post 2016 (4) Any BO fund with committed capital below $15 Million and any VC fund with committed capital below $10 Million. (5) Finally, any fund that did not have data on our selected set of predictors.

4.3.1 Categorical Variables

Most algorithms work with numerical data types in a feature vector. So, for a categorical variable in our data-set like "Fund_Geography" that can take three possible values "EU",

"USA" and "ASIA" we need to transform them into three separate vector represented as

$$EU = [1, 0, 0]$$

$$USA = [0, 1, 0]$$

$$ASIA = [0, 0, 1]$$

In addition to the above strategy, we have also employed our own custom one-hot encoding function that takes top 'x' number of labels for which one wants to hot encode, keeping rest as others. For instance categorical variable Fund_Type takes values such as Buyout, Venture, Funds of Funds, Early stage, Growth, Secondaries, Balanced, Expansion/Late stage up to 14 different categories in the raw data-set. We chose to one-hot encode it into Buyout and Venture cap dataset. This method is optimum for cases in which there are several categories with diminishing numerical representation. This method also helps avoid including noise into the data.

4.3.2 Feature Scaling

The next step in data processing involves normalizing the features to improve the efficiency of the models. With a few exceptions, ML algorithms do not perform well when its numerical features have different scales. While feature scaling is not a requirement for running all ML models, its application leads to faster convergence thus increasing the speed of learning and saving on computation expense. Additionally, scaling avoids the problem of numerical overflow while working with very small or very larger numbers as with their increasing values one needs more space to store it. There are two common methods used for resolving this issue: *Rescaling* and *Standardization*.

Rescaling (min-max scaling)

It is one of the simplest ways to normalizing numerical vectors that converts the feature into a standard range of values, typically in the interval $[-1, 1]$ or $[0, 1]$. This involves subtracting by the minimum value from the observed value and then divided it by the difference between the maximum and minimum value.

$$\bar{x}^{(j)} = \frac{x^{(j)} - min^{(j)}}{max^{(j)} - min^{(j)}}$$

where $min^{(j)}$ and $max^{(j)}$ are the minimum and maximum values of the feature j respectively.

Standardization

Standardization (z-score normalization) is the procedure in which feature values are re-scaled into a standard normal distribution. This is done by subtracting the mean value (so standardized values always have a zero mean) from the observation, and then divide it by the standard deviation - so that the resulting distribution has unit variance. Unlike min-max scaling, standardization does not bound values to a specific range, which may be a problem for some algorithms (e.g., neural networks often expect an input value ranging from 0 to 1). However, standardization is much less affected by outliers.

$$\hat{x}^{(j)} = \frac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}}$$

where $\mu^{(j)}$ and $\sigma^{(j)}$ are the mean (the average of the feature) and standard-deviation of feature j respectively.

The decision of which of the two methods is more suitable would depend on the feature and data-set in question. Standardization is usually preferred if the value of the feature takes a distribution that is close to a normal distribution or if the feature has outliers, since normalization would squeeze the outliers into a very small range.

Chapter 5

Methodology

5.1 Data Sampling

5.1.1 Train- Test Split

Once the data set is ready and annotated, the first set in any supervised machine learning (ML) algorithm is to divide the data into three subsets: Training set, Validation set, and Test set. The training set takes the bulk of the observations as it is used to build the model. The other two sets, often called holdout sets, are used to get an "out-of-sample" performance of the model. The is no defined optimal split percentage for ML, however, in general practice for smaller data like ours, the typical division is set at 70% training, 15% validation and 15% test. In practice, the focus of the analysis falls on the results obtained on test set, since the goal of any machine learning model is to accurately predict values for new data - that is data not used to train the model.

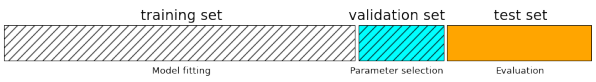


Figure 5.1: Data Sampling

5.1.2 Cross-Validation

It is a type of re-sampling method that involves repeatedly drawing samples from a training set and refitting the model to each of the samples in order to obtain additional information about the fitted model. Through this continuous process the model performance is assessed. So, we hold out a few samples or a fold in the training data, train our

model on the rest of the data and then validate on the hold out set. Cross-validation can be further divided into **Leave-p-out** cross-validation and **k-fold** cross validation.

We have used k-fold cross validation in our analysis that splits the data into **k** folds to validate the model on fold while training the model on the **k - 1** remaining folds, for **k** times.

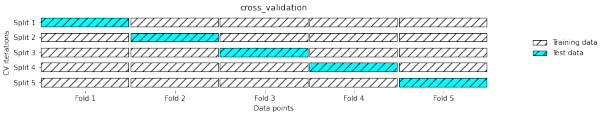


Figure 5.2: Cross validation

The error is then averaged over the k-folds.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} I(y_i \neq \hat{y}_i)$$

where I is an indicator variable

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \qquad .....(3) \\ 0 & \text{if } y_i = \hat{y}_i \end{cases}$$

Please note that the above loss function is also called Zero-One (0-1) loss function in which 1's become indicators for misclassified items. For instance,if we get two 1's from the function after evaluating 10 new samples, then the accuracy of the model is 80%.

Furthermore, we have used stratified k-fold cross validation that splits the data such that the proportion between classes that is 1,0/PME>=1, PME<1 are the same in each fold as they are in the whole dataset.

5.2 Machine Learning Models

5.2.1 General Framework

We start by defining the problem in matrix form, and then elaborating further depending on how each model deals with it. The general structure is to define the objective function

and then introduce parameterization penalties where ever necessary. Our aim for the rest of this chapter is to provide an in-depth description of each models so the reader does not require consulting any outside resource. We, however, do not elaborated on the computational techniques as its implementation can vary from one library to another.

Suppose we have a list of n training examples in the form:

$$\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$$

where $x^{(i)}$ is a vector of all the feature values(excluding the label) of the i^{th} instance in the dataset, and $y^{(i)}$ is its label(that is the target for the same instance).

X is a (nXp) matrix that denotes the space of input values , $X \in R^{nXp}$ that contains all the feature values (excluding labels) of all instance $1 : n$.

$$X = \begin{bmatrix} \left(x^{(1)}\right)^T \\ \left(x^{(2)}\right)^T \\ . \\ . \\ . \\ \left(x^{(n)}\right)^T \end{bmatrix} \qquad ...(4)$$

with $'n'$ as the number of observations(rows), and $'p'$ as the number of features/Predictors(columns).

$y^{(i)}$ is the output label that can be continuous or categorical. In our case is it categorical variable as our problem falls under the binary classification task, $y^{(i)} \in \{0, 1\}$. However, in general Y is the space of output values and is represented as

$$Y = \begin{bmatrix} \left(y^{(1)}\right) \\ \left(y^{(2)}\right) \\ . \\ . \\ . \\ \left(y^{(n)}\right) \end{bmatrix} \qquad ...(5)$$

Each row of the $X, (nXp)$ matrix is also referred to as **feature vector** of $'p'$ dimensions that is we define it as $\{(x^{(i)} \in R^p$ and always represent it as a column vector in isolation.

We represent each element in a feature vector as x_j^i where $'i'$ maps from $1 : n$ and j maps from $1 : p$.

5.2.2 Logistic Regression

Linear classifiers are set of models in which the decision boundary is a linear function of the input, unlike linear regression models, where the output $\hat{y}$ is a linear function of the features. As per the dimensions of the feature space, it can be a Line, Plane or a Hyper-plane. In a p dimensional space, a hyper-plane is a flat affine subspace of $p - 1$ dimensions. For instance for two predictors or features x_1 and x_2, a hyper-plane will be of one-dimension that is a line. The aim is to find fit the hyper-plane into the training data such that the division is optimal, and our prediction is maximised for new unseen sample.

To perform any supervised learning algorithm, we must decide on how to represent the hypothesis h that maps from $X \mapsto Y$ from (3) and (4) so that $h(x)$ is a good predictor for the corresponding value of y. In linear regression we approximate y as a linear function of x, so that we can write

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

where $\theta_i's$ are the parameters(weights) parametrizing the space of linear functions mapping from X to Y. We also assume $x_0 = 1$ so that we can represent the hypothesis as

$$h_\theta(x) = \sum_{i=0}^{p} \theta_i x_i = \theta^T x \qquad ...(6)$$

where p is the number of features(predictor variables) excluding x_0.

However, for a binary response with a 0/1 output as define in our case, the hypothesis function behaves poorly as some of the estimates from a linear regression might fall outside the [0,1] interval making them hard to interpret. Logistic regression solves this problem by not modelling Y directly but instead modeling the probability that Y belongs to a particular category.

This is done by tweaking the hypothesis function of the linear regression through a logistic function (also called sigmoid function) so that we write our new hypothesis as

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} \qquad ...(7)$$

where

$$g(z) = \frac{1}{1 + e^{-z}} \qquad ...(8)$$

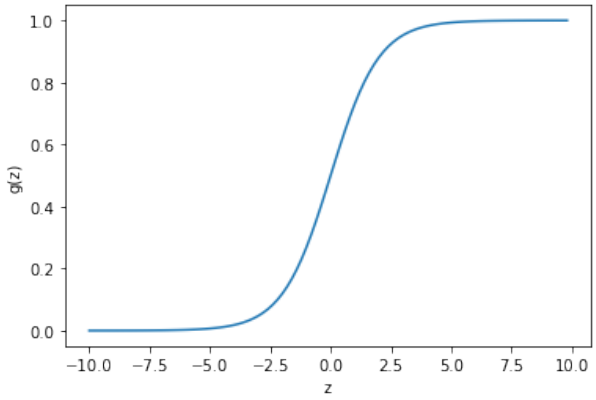


Figure 5.3: Sigmoid Function

Notice fig 5.5 as z approaches infinity $g(z)$ tends towards 1, and as z approaches negative infinity $g(z)$ tends towards 0, an ideal case to bound z between 0 and 1. Hence, $h_\theta(x)$ is also bounded between 0 and 1. Furthermore, say for threshold $g(z) = 0.5$, we can say

Predict 1, if $\theta^T x \geq 0 \rightarrow h_\theta(x) > 0.5$
Predict 0, if $\theta^T x \leq 0 \rightarrow h_\theta(x) < 0.5$

Given the training data set, we want to learn the parameters θ that makes $h_\theta(x)$ as close as possible to y. Formally, this function that measures for each value $\theta's$, how close the $h(x^{(i)})$'s are to the corresponding $y^{(i)}$ is called as cost function.

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} Cost(h_\theta(x^{(i)}), y^{(i)}) \qquad ...(9)$$

where

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = \begin{cases} -log(h_\theta(x^{(i)})) & \text{if } y^{(i)} = 1 \\ -log(1 - h_\theta(x^{(i)})) & \text{if } y^{(i)} = 0 \end{cases} \qquad ...(10)$$

The intuition behind equation (10) is that $Cost = 0$, if the true value $y^{(i)} = 1$, and through our hypothesis we predicted $h_\theta(x^{(i)}) = 1$, on the other hand the $Cost = -log(h_\theta(x^{(i)}))$, a very large value, if we predicted incorrectly i.e. $h_\theta(x^{(i)}) = 0$.

We can also say that as our prediction tends towards the incorrect zone, the cost increases and the above case can be represented as $h_\theta(x^{(i)}) \mapsto 0, Cost \mapsto \infty$

Similarly, if the true value is $y^{(i)} = 0$, and we predicted $h_\theta(x^{(i)}) = 1$, the cost will be $-log(1 - h_\theta(x^{(i)}))$ and zero for correct prediction.

The probabilistic view captures the same intuition that is cost will be zero if $h_\theta(x^{(i)}) = 0$, $P(y = 1|x; \theta) = 0$ read as (probability of y=1 given x ; parametrized by theta) however

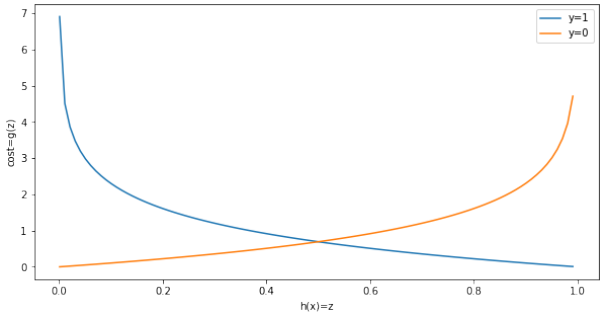


Figure 5.4: Logistic Regression Cost Function

as the probability will approach 1, the function will penalize with a very large cost. The above cost function can be defined from statistics using the principle of maximum likelihood estimation, which is an idea in statistics for how to efficiently find parameters' data for different models. We will not delve further as it is out of scope.

Equation (10) can also be written compactly as

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = -y^{(i)}.log(h_\theta(x^{(i)})) - (1 - y^{(i)}).log(1 - h_\theta(x^{(i)})) \qquad ...(11)$$

Now, if we sum the individual cost for our n samples then we define our objective function(also called as Loss function). Many authors use the terms Loss, Cost, Error, and Objective functions interchangeably. To clarify, we use Cost function for one training example and Loss or Objective function for the average of the cost function across all the examples.

Logistic regression loss function

$$J(\theta) = -\frac{1}{n}[\sum_{i=1}^{n} y^{(i)}.log(h_\theta(x^{(i)})) + (1 - y^{(i)}).log(1 - h_\theta(x^{(i)}))] \qquad ...(12)$$

Optimization: To fit the parameters θs we minimize the loss function through a gradient descent algorithm, which is a technique that basically finds the value of θs at which the loss function is minimum.

$$min_\theta J(\theta)$$

There are several other optimization algorithms for finding θs such as Conjugate gradient, BGFS, and L-GBFS. We have only tried implementing gradient descent. Standard available libraries bring the options of using other algorithms as well.

Gradient descent is based on the following rule

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \qquad ...(13)$$

where α is the learning hyper-parameter provided by the analyst. It takes partial derivative of J with respect to θ (the slope of J), and updates θ via each iteration with a selected learning rate α until the Gradient Descent has converged.

As discussed earlier if we have too many features, the learned hypothesis $h_\theta(x)$ may fit very well on the training data set by minimizing the loss function $J(\theta)$ but fail to have a good out of sample performance. This problem is defined as Overfitting, resulting in a complex with a low bias but high variance. To tackle overfitting, one can employ the following three methods - Subset Selection, Shrinkage methods and Principal component analysis. In our analysis, we have employed the shrinkage method.

Subset Selection: It is an approach that involves identifying a subset. We can do this by reducing the number of features - i.e. manually selecting a few that we believe is related to the response. The ideal candidate(metric) for numerical features is the *Pearson correlation Coefficient*. It must be noted that Pearson correlation coefficient cannot work on categorical variables, therefore, one has to hot encode the categorical variables before finding the correlations amongst the features. A few papers also suggest to use *Cramer's V* in place of Pearson coefficient for better visualization of the correlation among the categorical variables.

Step-wise Selection: To safely select features that have high correlation with the target variable we can use with one of the two set-ups in Step-wise Selection namely Forward step-wise selection and Backward step-wise selection. In forward step-wise selection we begin with a model containing no predictors, and then keep on adding predictors one-at-time, until all the predictors are in the model. At each step we add a feature that

gives the highest additional improvement against some metric say Precision in our case or OLS in regression. Contrarily, in Backward selection, we start with all predictors and iteratively remove the least useful predictor, one at a time. Please note that we have not applied any Sub-selection procedure in choosing the number of predictors.

Shrinkage Method : Unlike subset selection, in which we choose a few variables among all p variables, in the Shrinkage method, we constrain or regularize the co-efficient estimates towards zero. To do this we add penalty terms in our loss function. So equation (12) becomes

in case of **Ridge** (also called **l2 norm regularization**)

$$J(\theta) = -\frac{1}{n}[\sum_{i=1}^{n} y^{(i)}.log(h_\theta(x^{(i)})) + (1 - y^{(i)}).log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2n}\sum_{j=1}^{p}\theta_j^2 \qquad ...(14)$$

and in case of **Lasso** (also called **l1 norm regularization**)

$$J(\theta) = -\frac{1}{n}[\sum_{i=1}^{n} y^{(i)}.log(h_\theta(x^{(i)})) + (1 - y^{(i)}).log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2n}\sum_{j=1}^{p} \mid \theta_j \mid \qquad ...(15)$$

where $\lambda \geq 0$ is called as the tuning parameter. When λ is 0 then the penalty term has no effect, however as it approaches ∞ then the impact grows, so that the coefficients or the θ_j's approach zero. Notice that if we use a very high value of λ then we shrink all the co-efficients to zero thereby shrinking the model to its intercept. Lasso penalty,like ridge, also shrinks the coefficient estimates towards zero, however, it has the effect of forcing a few of the parameters to be exactly zero. This entails that Lasso penalty also performs variable selection as done by the class of methods in Subset selection.

The third type of regularizer is called **Elastic Net** that uses both l1-norm and l2-norm regularization in some sort of trade off manner controlled by the hyper-parameter ρ. The elastic net penalty is represented as

$$J(\theta) = -\frac{1}{n}[\sum_{i=1}^{n} y^{(i)}.log(h_\theta(x^{(i)}))+(1-y^{(i)}).log(1-h_\theta(x^{(i)}))]+\frac{\lambda}{2n}\sum_{j=1}^{p}[\rho\theta_j^2+(1-\rho).\mid\theta_j\mid] \quad ...(16)$$

We demonstrate the effect on the precision by using k-fold cross-validation and regularization in our logistic regression model.

5.2.3 Discriminant Analysis

Discriminant analysis encompasses methods that can be used either for Classification or dimensionality reduction. Since, our topic at hand is binary classification, we will not deal with the dimensionality reduction aspect of it. Broadly, it can be divided into Linear Discriminant analysis (LDA), Quadratic discriminant analysis (QDA) and a compromise between two, that is called as Regularized Discriminant Analysis. Our focus is only on the application of LDA and QDA, and chart out how these algorithms perform in comparison to the logistic regression and the Bayesian error rate (that is the minimum classification error rate).

logistic regression involves direct modelling of the input space to the output space using the logistic(sigmoid) function and is represented as $Pr(Y = k|X = x;\theta)$. This representation is basically finding the probability of the output label Y to be in a class k given the predictors X and parametrized by θ. We can also call say that the model is the conditional distribution of the response Y given the predictors X.

The alternative approach and less direct approach to estimate such probabilities is by using the Bayes' theorem. So, we model the distribution of the predictors X separately in each of the response classes Y, and then use Bayes' theorem to find the probability.

We are not using the Bayes' classification algorithm but a brief introduction to the theorem is warranted as both LDA and QDA are based on it. Bayes' theorem states that

Like others, **LDA** is also a supervised learning technique, that classifies with a linear decision boundary, generated by fitting class conditional densities and using Bayes' rule. The model assumes that $X = (X_1, X_2, ..., X_p)$ is drawn from a multi-variate Gaussian distribution with a class specific mean vector and a common co-variance matrix.

For p-dimensional random variable X (as the number of predictors is 'p') with a multi-variate Gaussian distribution, we represent $X \sim \mathrm{N}(\mu_k, \sum)$, where μ_k is the class specific vector and $\sum$ is the co-variance matrix common to all classes.

The discriminant function for LDA takes a linear form is defined as

$$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2} \mu_k^T \sum^{-1} \mu_k + log\pi_k$$

It is important to note here that despite assuming that random samples are drawn from multi-variate Gaussian distribution, we still need to estimate parameters $\mu_1, \mu_2, ...\mu_k$ and $\pi_1, \pi_2, ...\pi_k$ for each class $k \in 1, 2, ..K$.

Similarly, **QDA** also assumes that observations from each class are drawn from a Gaussian distribution, however, unlike LDA, QDA assumes that each class has its own co-variance matrix. Therefore an observation of k_{th} class is represented as $X \sim \mathrm{N}(\mu_k, \sum_k)$, where $\sum_k$ is a co-variance matrix of the k_{th} class.

The discriminant function in QDA takes a quadratic form and is represented as

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T . \sum_k^{-1} (x - \mu_k) - \frac{1}{2} log \mid \sum_k \mid + log\pi_k$$

The other difference between LDA and QDA also emanates from the bias-variance trade off. For 'p' predictors, estimating a co-variance matrix means estimating $\frac{p(p+1)}{2}$ parameters that is the case in LDA, however, in QDA when we assume different co-variance matrix for each class then we have to estimate $\frac{k.p(p+1)}{2}$ parameters. So, LDA with fewer parameters is a much less flexible classifier than QDA, and so has lower variance, i.e. higher bias in case the same co-variance matrix assumption doesn't fit the actual(ground truth) distribution.

Overall, LDA tends to perform better than QDA when we have fewer training observations as reducing variance is important. For data sets with large training data set, QDA is recommended.

5.2.4 Support Vector Classifier

Support Vector Classifier(SVC) is another linear classifier that we have tried on our data and bears some parallels with the logistic regression. Support Vector Machine(SVM) is an extension of SVC that uses a kernel trick for non-linear class boundaries. SVC is also sometimes called as SVM without kernels or Linear SVM.

As discussed in the logistic regression that tries to fit the best decision boundary for the classification problem, SVM also tries to fit a separating hyper-plane that separates the training observations perfectly according to their class labels. Now, if our data is linearly separable then we can have infinite numbers of hyper-planes. However, out of this infinite possibilities if we choose a hyper-plane that not only separates the two classes(say our case of binary classification) but also stays away from the closest training instances as possible. In other words one can think it as fitting the widest possible street or slab(represented by parallel lines) between the classes. The middle line of this widest street or slab is called as **maximal margin hyper-plane**(also known as optimal separating hyper-plane). Maximal margin hyper-plane is the separating hyper-plane for which the margin is largest.

The training instances or examples that lie on the two parallel dashed lines are called as Support vectors as they are represented in p dimension space. Notice in fig 5.5 that adding more examples off the street/slab will not effect the decision boundary at all, as the decision boundary is based only on the support vectors.

The equation of such a hyper-plane can be defined as $\theta^T x = 0$ and we segregate classes that is predict 1 if $\theta^T x > 0$, otherwise 0. So we are fitting our model to find our parameters θs such that it gives the best prediction or say the best linear classification.

We define the hypothesis of the SVM as

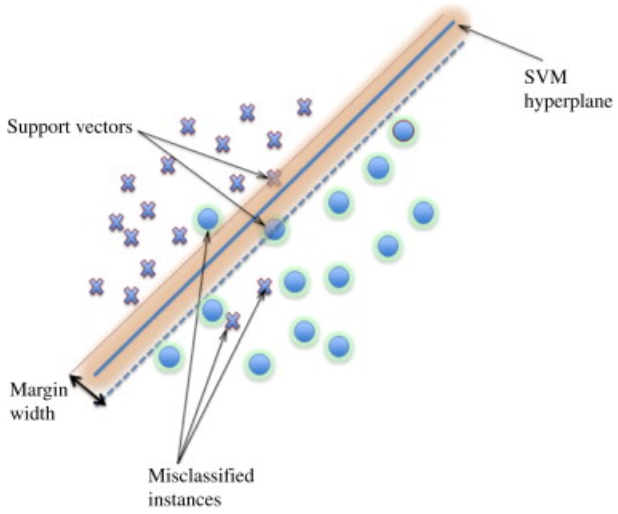


Figure 5.5: SVM image display

$$h_\theta(x) = \begin{cases} 1 & \text{if } \theta^T x \qquad .....(17) \\ 0 & otherwise \end{cases}$$

SVM uses hinge loss as its cost function. The intuition is that the cost will be zero if actual value $y^{(i)}$ is 1, and our prediction $\theta^T x \geq 1$. The cost increases as the value of the $\theta^T x$ becomes less than 1.

Similarly, the cost will be zero if the actual output $y^{(i)}$ is 0, and we predicted $\theta^T x \leq -1$. Cost will increase if our prediction becomes greater than -1.

One can notice that SVM punishes both incorrect prediction as well as those that are within the street/slab. We can also conclude that $\theta^x = 1$ and $\theta^x = -1$ are the equations of the boundaries that hold the support vectors. This also means the SVM's decision boundary doesn't depends on non support-vectors, if we change or remove them - the total value of the cost function won't change.

We can write the cost function using the hinge loss function.

$$Cost(h_\theta(x^{(i)}), y^{(i)}) = \begin{cases} max(0, 1 - \theta^T x) & \text{if } y^{(i)} = 1 \qquad ...(18) \\ max(0, 1 + \theta^T x) & \text{if } y^{(i)} = 0 \end{cases}$$

SVC objective function

$$J(\theta) = C[\sum_{i=1}^{n} y^{(i)}.max(0, 1 - \theta^T x) + (1 - y^{(i)}).max(0, 1 + \theta^T x)] + \frac{1}{2n}\sum_{j=1}^{p} \theta_j^2 \qquad ...(19)$$

where C is the tuning parameter that plays a similar role to $\frac{1}{\lambda}$. To reiterate, both C and λ prioritize how much we care about optimize fit term and regularized term. The observant reader will notice that we have added the ridge regularization penalty in equation (19). C determines the severity of the violations to the margin(and to the hyperplane) that we can tolerate as the value of C is provided by the analyst. For increasing value of C, the width of the street/slab decreases, and this is only possible when we have clear separability of the two classes. In other words narrow margin means rare violation, highly fitting model, low bias and high variance. As C decreases, the margin(width of the street) becomes large and we become more tolerant towards incorrect classifications. In other words broader margin means more violations, less fitting model, high bias and low variance.

We observe that linear SVM and Logistic regression behave similarly with comparable Precision values as the logistic loss and hinge loss are comparable. However, if we change our SVM objective function to tackle non-linear decision boundaries then we can see marginal improvement.

5.2.5 K-Nearest Neighbours

Parametric methods such as logistic regression assumes a linear function form for the model, and sometimes suffers from the strong assumptions about the form of $f(X)$. If the assumption itself is far from the ground truth then such models will perform badly.

In contrast, non-parametric methods such as KNN do not assume parametric from of $f(X)$ rather classify by estimating the conditional distribution of Y given X, and alloting it to the class with the highest estimated probability.

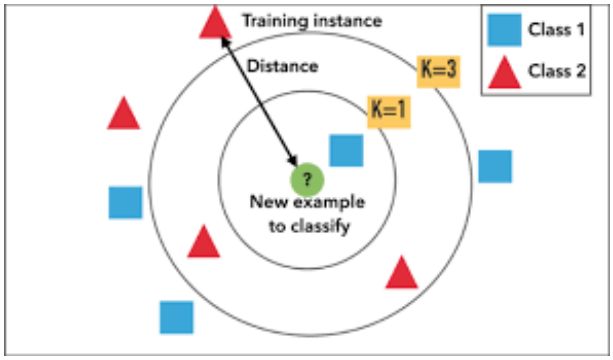


Figure 5.6: Nearest Neighbour

Given a positive integer k and a test observation x_0, the KNN classifier first identifies the k points in the training data that are closest to x_0, represented as N_o. It then estimates the conditional probability for class j as the function of points in N_o whose response values equal j.

$$Pr(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N_o} I(y_i = j) \qquad ...(21)$$

Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability. The choice of K has a drastic effect on the KNN classifier obtained. That is as K grows, the method becomes less flexible and produces a decision boundary that is close to linear, a low-variance and high-bias classifier

5.2.6 Multi-layer Perceptron Model

Multi-layer perceptron model learns a non-linear function $f : R^p \mapsto R^o$ that maps from p dimensional feature space to an output space depending on the desired problem at hand. For examples, output space O will be 1 for a regression task and 2 for binary classification, that is our case. The first layer of the model is called as the input layers and is represented as

$$\{x_j : |x_1, x_2, ..x_p\}$$

Each neuron or unit in the hidden layer transforms the values from the previous layer

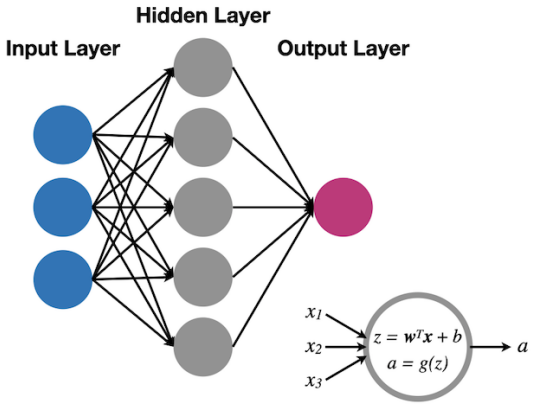


Figure 5.7: Neural Network

with a weighted linear summation represented as

$$w_1x_1 + w_2x_2 + ... + w_px_p$$

where $w_1, w_2, ...w_p$ are the weights. This is transformation is then followed by a non-linear activation function $g(.)$ such as TanH, RELU and Sigmoid function. It is important to note different activation functions and how they differentiate among themselves. Additionally, the choice of an activation function is the discretion of the researcher. Finally the output layer receives the values from the last hidden layer and transforms them into output values.

5.3 Comparison Measures

After having defined the set of models we are using in our analysis, the next step is to select the metrics we are going to use to compare them. The chosen analytical approach depends on the business dynamics or the business use-case at hand. For any classification problem, the default method is to start with the analysis of the confusion matrix.

5.3.1 Precision, Accuracy & Recall

The two most frequently used metrics to assess any classification model are **Precision** and **Recall**. Precision is the ratio of correct positive predictions to the overall number of

positive predictions, whereas Recall is the ratio of correct positive predictions to the overall number of positive examples in the dataset. A typical confusion matrix is represented as follows.

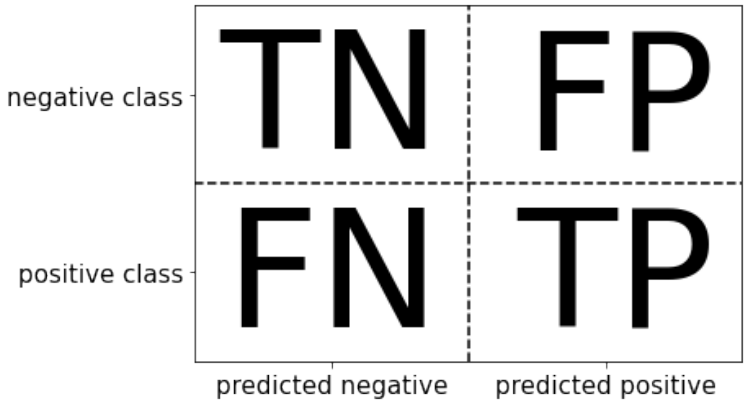


Figure 5.8: Confusion Matrix

Formulas:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Almost always, in practice, we have to choose between a high precision or a high recall. It is usually impossible to have both. In our hypothesis, in which we want to maximize cases with PME greater than 1, Precision is more important than recall. Precision will return the proportion of correct use cases in the list of all returned use cases. In other words, a precision of 80% means that out of 100, our investment will be precise only in 80 of them, with rest going as False Positives. On the other hand, Recall is the ratio of relevant use cases to the total number of relevant use cases that could have been returned.

So, a recall of 45% means that our investment thesis will be correct only for 45 out of 100 investment cases that we are going to do in any case because we predicted in total 100 cases as Positive i.e. with PME >= 1.

So, having higher precision means that we will have higher proportion of correct investment cases i.e. we are avoiding mistakes by detecting use cases with less than 1 as legitimate, however we are ready to tolerate lower recall with some use cases that we will be not investing. Higher precision means less loss as we will invest in all use cases that we have predicted PME>1. Bottom line is that we would tolerate False positive less than False negatives. False negatives can be high after all we are not going to invest in those cases.

Overall, the ideal situation for a confusion matrix should be that our prediction is only divided into True Positives and True Negatives. Practically, such case is impossible, and we have to choose between False Negatives and False Positive, that is selecting one that is going to hurt less. Is it higher False Positive or lower Precision or is it higher False Negative or lower Recall that is going to hurt depends on the business case at hand. In our case, it is lower Precision that will hurt LPs investment thesis as LP can tolerate lower recall by loosing a few investments.

5.3.2 AUC - ROC Curve

The other metric that we employed for model comparison is **AUC (Area Under the Curve)** in a ROC (Receiver Operating Characteristics) graph. A ROC curve is a graphical plot that summarizes the trade-off between the true positive rate (Recall) and false positive rate for a predictive model at different probability thresholds, thus representing the capability of the classifier in distinguishing the the two classes. To compare models, AUC is used to summarizes the performance of each classifier into a single measure by calculating the area under the ROC curve. The bigger the area under the curve, the better is the classifier. The metric was borrowed from the signal detection theory but has now has gained importance in several fields including machine learning - particularly in

the case of classification problems.

Sometimes, the ROC curve is also defined in terms of Specificity and Sensitivity. Note that - Sensitivity, Recall and True positive rate are all different names for the same ratio. To set some context for a reader familiar with the terms used in other domains, we provide few related formulas.

- Sensitivity = True positive rate = Recall = TP/(TP+FN))
- Specificity = True negative rate = selectivity = TN/(FP+TN)
- 1 – Specificity = False Positive rate = FP/(FP+TN))

Chapter 6

Analysis and Results

We start this chapter by comparing our calibrated models first though their levels of "out of sample[1]" accuracy and precision and then complementing the numbers using the AUC-ROC curve values. Given the differentiated investment style of Buyout and Venture Capital funds, we ran separate models for each fund type. We then benchmark the top predictions of two of our models - Logistic and LDA, against a naive investment strategy. We conclude this chapter by offering some insights into the features driving the predictive power of these models and compare these findings to past research.

6.1 Model Comparison

The confusion matrix for all the models in this section are based on a probability threshold of 0.5, which means that if the estimates probability of a fund beating the performance threshold (PME > 1) is over 50%, then the target would take the value 1 and if they probability is lower than 50%, then the target would take the value 0. In practice, we can adjust this threshold anywhere between 0 and 1 depending on the level of certainty the user is looking for.

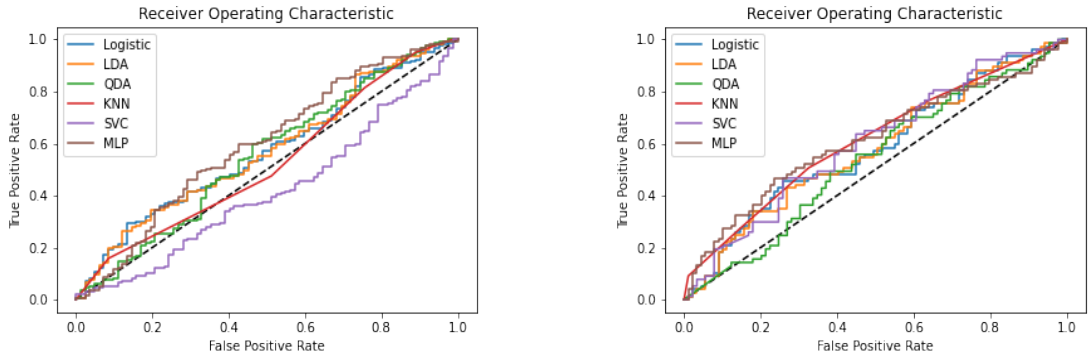
Based on the accuracy numbers presented in Table 6.1, we observe that Logistic, LDA and SVM show the most promising results for buyout (BO) funds with a top accuracy of 69%. The leveled performance of these three models can be attributed to the similarity in their underlying framework that tries to fit a linear decision boundary among the observations. This further suggests that the relationships between variables is possibility linear for BO funds. The level of precision is similar across all models at this probability threshold. On the other hand, for venture capital (VC) funds the nearest neighbor (KNN)

[1]Out of Sample refers to the values we get for the "**test**" set

Models	Buyout			Venture Capital		
	Accuracy	Precision	AUC	Accuracy	Precision	AUC
Logistic Regression	0.69	0.70	0.56	0.54	0.52	0.58
Linear Discriminant Analysis (LDA)	0.69	0.70	0.56	0.53	0.50	0.58
Quadratic Discriminant Analysis (QDA)	0.66	0.71	0.54	0.50	0.47	0.52
Support Vector Classifier (SVC)	0.69	0.70	0.51	0.56	0.57	0.59
Nearest Neighbour (KNN)	0.63	0.70	0.52	0.61	0.58	0.57
Neural Network (MLP)	0.62	0.71	0.56	0.60	0.57	0.59

Table 6.1: Summary of Model Performance (Out of Sample)

and neural network (MLP) outperform the more basic models both in terms of accuracy (60%) and precision (58%) . Our preliminary analysis of the ROC curves in figure 6.1, follow suit with our results from the accuracy comparison with the Logistic and LDA models performing the best for BO. Results differ slightly for VC funds, where only QDA seems to perform poorly relative to the other models. Over-all, none of our models have an AUC below 0.5, reflecting at least some superiority in predicting results over random selection.



(a) Buyout

(b) Venture Cap

Figure 6.1: AUC - ROC Comparison

6.2 Naïve vs Machine Learning Strategy

To get a real-world estimate of the quality of our predictions, we compare a set of performance measures (Net IRR, Net Multiple and PME) for a portfolio of funds that our model predicts is most likely to beat the market - that is cross the PME threshold of 1, against a portfolio of funds proposed by a naïve investment strategy. We use the Logistic

and LDA models for both BO and VC funds to produced two portfolios of funds as the machine learning benchmark. These are the top 10% funds out of the test sample that have the highest probability of beating the performance threshold (PME > 1). As for the "naïve strategy" – we invest in the biggest funds in the sample by choosing the top 10% funds by committed capital - going by a strategy of targeting funds by size. For simplicity, we assume that equal amount of capital is invested in each fund for all three strategies.

Tables 6.2 & 6.3 present the performance of each portfolio of funds in terms of average Net IRR, average Net Multiple and average PME along with their corresponding standard deviations. For BO funds, where the portfolio size is 26 funds, we observe that the naïve strategy portfolio performs better than our models in terms of net IRR but has a lower net multiple and PME. A possible explanation for this could be that our models are calibrated for filleting out funds based on PME and not Net IRR. For VC funds, where each portfolio has 16 funds, both of our ML models outperform the naïve strategy portfolio in terms of net IRR and multiple and are only marginally lower in terms of average PME. It is also interesting to note that stand deviation of the ML portfolios are lower than the naïve strategy portfolio particularly in the case of Net IRR, indicating consistency in portfolio returns.

Table 6.2: Portfolio Performance Summary (BO)

Model	Avg Net IRR	stdev	Avg Net Multiple	stdev	Avg PME	stdev
Logistic	13.12	11.10	1.70	0.44	1.21	0.35
LDA	13.83	10.11	1.73	0.41	1.24	0.32
Naive Stratergy (Size)	14.97	9.49	1.57	0.28	1.13	0.18

Table 6.3: Portfolio Performance Summary (VC)

Model	Avg Net IRR	stdev	Avg Net Multiple	stdev	Avg PME	stdev
Logistic	11.68	10.32	1.53	0.60	1.03	0.36
LDA	9.98	10.95	1.51	0.65	1.00	0.42
Naive Stratergy (Size)	8.61	15.26	1.43	0.65	1.13	0.44

6.3 Analysis of Predictors

While the primary aim of most prediction models like ours is to focus on the quality of the forecasts - be it in terms of accuracy or precision; being able to interpret the factors driving these results is also necessary to validate and improve on our work. The more advanced models like neutral network are often considered "black boxes" and are criticized for their superior predictability coming at the cost of interpretability. Interestingly, for our data the more basic models like logistic regression are performing almost at par, if not better than advanced models such as neural networks, potentially owing to the limited size sample. We take a closer look at results from our logistic regression and summarize our findings for the variables that are driving the model's predictive power, comparing them to empirical findings wherever possible. For avoiding the results to be skewed by the sample split we do for the training and test set, we train the model using 99% of the BO and VC funds samples.

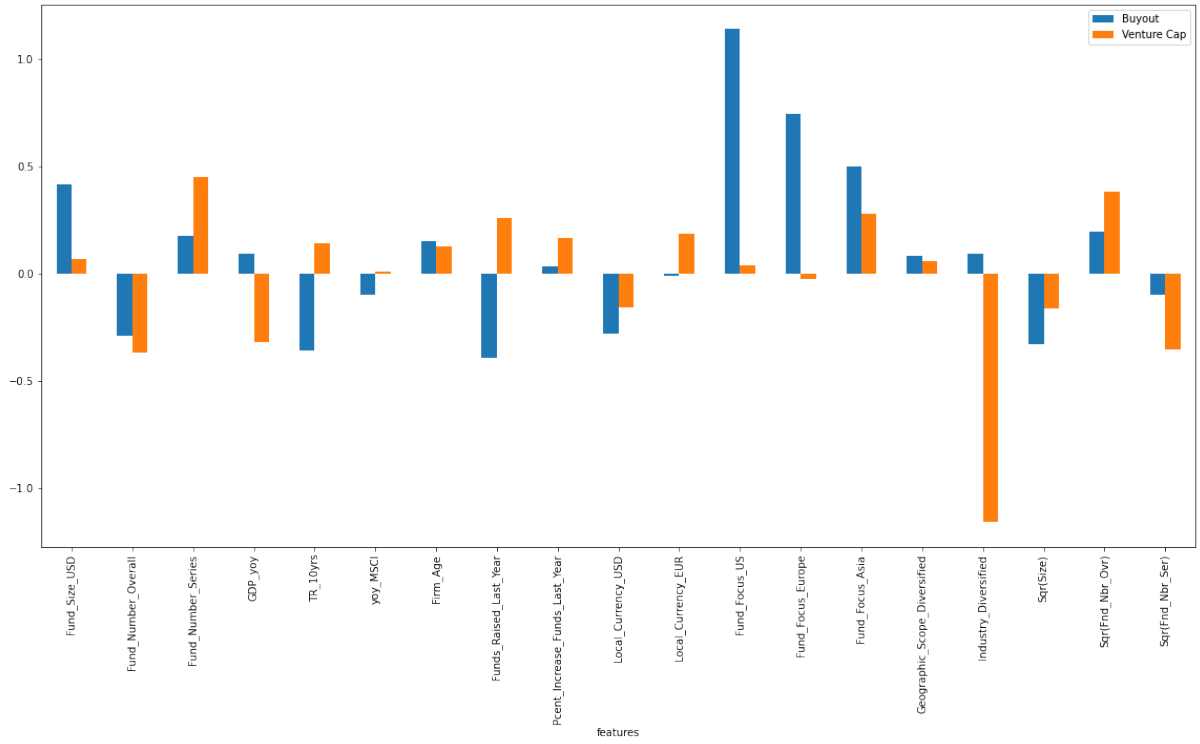


Figure 6.2: Coefficients of Predictors (BO & VC)

Consistent with the findings of Kaplan and Schoar (2005) and Gottschalg et al. (2004), our logistic model also observes a concave relationship between fund size and fund per-

formance for both BO and VC funds - that is a positive coefficient for fund size but a negative one for its squared term. The effect is stronger on BO funds, which is justified by the higher variance we observe in their targeted fund size. The management experience measured in terms of firm's age (how long the GP has been in business and the fund's sequence number (the chronological number of the fund raised by the GP) shows differentiated results. The firm age at the time of fund raising shows a positive effect on fund performance for both fund types. The fund's (overall) sequence number on performance however seems to have a positive effect only at very high sequence numbers for both VC and BO funds, in line with Roggi et al (2019)'s findings of a convex relationship between fund sequence number and performance. Interestingly, the fund series number shows a concave relationship instead, which could suggest that initial funds of a spinoff series perform well. We also observed marginal increase in the Accuracy and Precision after adding macro-economic variables such as the US GDP y-o-y growth and 10-year US treasury bond rate during the fundraising year, with opposing coefficients for BO and VC funds. High treasury bond yields during the time of fundraising shows a negative effect on BO fund performance which is understandable given their heavy reliance on debt financing. Industry specialization seems to be a key performance driver for VC funds, going against the finding of Lossen et al (2006) [20] however, BO fund returns still show some degree of benefit from diversifying across industries.

Chapter 7

Conclusion

In this chapter, we summarize the findings of our study and address the research questions we presented in the introduction chapter. We further discuss the limitations we faced in our research and offer recommendations on how future studies can improve on our findings.

7.1 Conclusion

Out study aimed to train a range of machine learning models into a binary classification setting aimed at determining the likelihood of a fund exceeding a pre-determined performance threshold using a set of parameters available to LPs during the fundraising year. Our target variable is constructed using the Kaplan-Schoar PME and takes the value **1**, if the fund exceed the PME value of one, and takes the value **0** if the PME is less than one. Our predictor variables consisted of a mix of *fund specific features* like - targeted fund size, management experience, industry and geographical specialization; as well as indicators of the *macroeconomic environment* at the time of investment selection such as: the fundraising year's GDP growth rate, MSCI World Index growth, current volume of funds raised and prevailing interest rates.

The results presented especially for the buyout (BO) funds make a convincing case for the ability of machine learning models in predicting the fund performance with the top models showing an accuracy of 69%, while the highest accuracy achieved for VC funds was 61% . Our findings show encouraging signs of machine learning's applicability for complementing limited partner's due diligence process. At the highest level, the best performing models for BO funds were Logistic Regression, Linear Discriminant Analysis and Support Vector Machines, clearly outlining the prevalence of the linear decision bound-

aries for our datasets. Shallow learning clearly outperformed deep learning - a result, which is atypical from other research areas, which we attribute to dearth of data. This was also one of the reasons we were unable to experiment too much with advance models like neural networks by adjusting the number of neurons and trying different activation functions. For VC funds, the non-parametric model – KNN showed the best results in terms of accuracy suggesting a more non-linear relationship between the variables.

The growing volume of PE funds demands a tool that could skim through hundreds of options simultaneously and allow LPs to focus only a subset of "quality funds" that offers the highest probability of success. Thus, employing AI and machine learning models like ours could be a highly complementary tool for LPs in their investment decision process.

7.2 Limitation & Future Research

While relying on past fund performance alone is a sub-optimal method for judging subsequent fund's performance, it could be an important variable when fitted alongside other predictors and explored through machine learning models. Due to the limitations of our data-set, a measure for past fund performance is missing from our model. It would be interesting to see if the addition of this variable in future studies could make a significant improvement to our model's predictive power. Another potential issue we face is the use of a large portion of non-liquidated funds in our analysis, which could lead to misleading results if the NAVs we use for calculating our PME values are biased due to the self-reporting by GPs. While we take several steps to correct for this issue, our dataset could still be affected by this problem. Training models with primarily liquidated funds could lead to more reliable results. Furthermore, our study only applies a small subset of information available to LPs during fundraising. It would be interesting to see if adding proprietary level data into the analysis would make a significant improvement in our results.

References

[1] Philipp Aigner et al. "What drives PE? Analyses of success factors for private equity funds". In: *Journal of Private Equity* 11.4 (2008), pp. 63–85. ISSN: 10965572. DOI: 10.3905/jpe.2008.710907.

[2] Bain & Company. *Bain & Company (2021) Global Private Equity Report*. URL: https://www.bain.com/globalassets/noindex/2021/bain%7B%5C_%7Dreport%7B%5C_%7D2021-global-private-equity-report.pdf (visited on 04/24/2021).

[3] Gregory W Brown, Oleg R Gredil, and Steven N Kaplan. "Do private equity funds manipulate reported returns?" In: *Journal of Financial Economics* 132.2 (2019), pp. 267–297. ISSN: 0304405X. DOI: 10.1016/j.jfineco.2018.10.011. URL: https://doi.org/10.1016/j.jfineco.2018.10.011.

[4] Gregory W. Brown et al. "What Do Different Commercial Data Sets Tell Us About Private Equity Performance?" In: *SSRN Electronic Journal* (2015). ISSN: 1556-5068. DOI: 10.2139/ssrn.2701317.

[5] Dan Cheng and Pasquale Cirillo. "A reinforced urn process modeling of recovery rates and recovery times". In: *Journal of Banking and Finance* 96 (Nov. 2018), pp. 1–17. ISSN: 03784266. DOI: 10.1016/j.jbankfin.2018.08.014.

[6] Michael Chui, Vishnu Kamalnath, and Brian McCarthy. "An executives guide to AI". In: *McKinsey Analytics* Feb (2018), p. 12. URL: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai.

[7] "Enhancing Private Equity Manager Selection with Deeper Data". In: September (2017).

[8] EQT. *Motherbrain - EQT*. URL: https://www.eqtgroup.com/digital/motherbrain/ (visited on 04/24/2021).

[9] Joachim Freyberger, Andreas Neuhierl, and Michael Weber. "Dissecting Characteristics Nonparametrically". In: *Review of Financial Studies* 33 (2020).

[10] Shihao Gu, Bryan Kelly, and Dacheng Xiu. "Empirical Asset Pricing via Machine Learning *". In: *The Review of Financial Studies* 33 (2020), pp. 2223–2273. DOI: 10.1093/rfs/hhaa009. URL: https://academic.oup.com/rfs/article-abstract/33/5/2223/5758276.

[11] Robert S Harris et al. "Has Persistence Persisted in Private Equity? Evidence from Buyout and Venture Capital Funds". In: *SSRN Electronic Journal* (2020). DOI: 10.2139/ssrn.3736098.

[12] Delvin D Hawley, John D Johnson, and Dijjotam Raina. "Artificial Neural Systems: A New Tool for Financial Decision-Making". In: *Financial Analysts Journal* 46.6 (1990), pp. 63–72. DOI: 10.2469/faj.v46.n6.63. URL: https://doi.org/10.2469/faj.v46.n6.63.

[13] Chris Higson and Rüdiger Stucke. "The Performance of Private Equity". In: *SSRN Electronic Journal* (Feb. 2012). URL: https://papers.ssrn.com/abstract=2009067.

[14] Tim Jenkinson, Miguel Sousa, and Rüdiger Stucke. "How Fair are the Valuations of Private Equity Funds?" In: *SSRN Electronic Journal* (Mar. 2013). ISSN: 1556-5068. DOI: 10.2139/ssrn.2229547. URL: https://papers.ssrn.com/abstract=2229547.

[15] Steven N. Kaplan and Antoinette Schoar. "Private equity performance: Returns, persistence, and capital flows". In: *Journal of Finance* 60.4 (2005), pp. 1791–1823. ISSN: 00221082. DOI: 10.1111/j.1540-6261.2005.00780.x.

[16] Steven N. Kaplan and Per Strömberg. "Leveraged buyouts and private equity". In: *Journal of Economic Perspectives* 23.1 (2009), pp. 121–146. ISSN: 08953309. DOI: 10.1257/jep.23.1.121.

[17] Bryan T. Kelly, Seth Pruitt, and Yinan Su. "Characteristics are covariances: A unified model of risk and return". In: *Journal of Financial Economics* 134.3 (Dec. 2019), pp. 501–524. ISSN: 0304405X. DOI: 10.1016/j.jfineco.2019.05.001. URL:

https://asu.pure.elsevier.com/en/publications/characteristics-are-covariances-a-unified-model-of-risk-and-retur.

[18] Amir E. Khandani et al. "Consumer credit-risk models via machine-learning algorithms". In: 34.11 (2010), pp. 2767–2787. URL: https://econpapers.repec.org/RePEc:eee:jbfina:v:34:y:2010:i:11:p:2767-2787.

[19] Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2 (2020), pp. 271–292.

[20] Ulrich L. "The Performance of Private Equity Funds: Does Diversification Matter? by Ulrich Lossen :: SSRN". In: 22.4 (2009), pp. 1747–1776.

[21] Austin Long and Craig Nickels. "A Private Investment Benchmark". In: *AIMR Conference on Venture Capital Investing* (1996), pp. 1–17.

[22] McKinsey. "A new decade for private markets: McKinsey Global Private Markets Review 2020". In: February (2020), pp. 1–42.

[23] Ignacio Olmeda and Eugenio Fernandez. "Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction". In: *Computational Economics* 10.4 (1997), pp. 317–35.

[24] Ludovic Phalippou and Oliver Gottschalg. "The Performance of Private Equity Funds Author (s): Ludovic Phalippou and Oliver Gottschalg Published by : Oxford University Press . Sponsor : The Society for Financial Studies . Stable URL : https://www.jstor.org/stable/30225708 The Performance of Priv". In: 22.4 (2009), pp. 1747–1776.

[25] Ludovic Phalippou and Maurizio Zollo. "What Drives Private Equity Fund Performance?" In: *Working Papers – Financial Institutions Center at The Wharton School* November (2006), pp. 1–29. ISSN: 00221082. URL: http://fic.wharton.upenn.edu/fic/papers/05/0541.pdf.

[26] Thomas Renault. "Intraday online investor sentiment and return patterns in the U.S. stock market". In: *Journal of Banking and Finance* 84 (Nov. 2017), pp. 25–40. ISSN: 03784266. DOI: 10.1016/j.jbankfin.2017.07.002.

[27] David T Robinson and Berk A Sensoy. "Cyclicality, performance measurement, and cash flow liquidity in private equity". In: *Journal of Financial Economics* 122.3 (2016), pp. 521–543. ISSN: 0304405X. DOI: 10.1016/j.jfineco.2016.09.008. URL: http://dx.doi.org/10.1016/j.jfineco.2016.09.008.

[28] Oliviero Roggi et al. "Private equity characteristics and performance: An analysis of North American venture capital and buyout funds". In: *Economic Notes* 48.2 (2019). ISSN: 14680300. DOI: 10.1111/ecno.12128.

[29] Morten Sorensen and Ravi Jagannathan. "The Public Market Equivalent and Private Equity Performance". In: *SSRN Electronic Journal* (2013). ISSN: 0015-198X. DOI: 10.2139/ssrn.2259261.

[30] Dong Zhao et al. "An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach". In: 49.2 (2017), pp. 325–341. URL: https://link.springer.com/article/10.1007/s10614-016-9562-7.

Appendix A

A.1 Summary Statistics (Vintage Year)

Table A.1: Buyout Funds

Vintage Year	Number of Funds	Average Size ($M)	Average Net IRR(%)	Average MOIC	Average PME
1980	1	60	32.1	11.87	3.84
1981	0	-	-	-	-
1982	0	-	-	-	-
1983	0	-	-	-	-
1984	0	-	-	-	-
1985	2	589	10.72	1.96	1.14
1986	1	59	34.4	2.94	1.67
1987	5	321.6	23.24	4.19	1.95
1988	4	322.25	15.92	2.01	1.26
1989	3	268.19	20.11	2.51	1.48
1990	4	281	21.94	2.62	1.41
1991	3	356	23.47	2.64	1.27
1992	6	108.25	12.79	2.07	1.12
1993	8	422.46	28.08	2.57	1.42
1994	16	464.64	23	2.06	1.32
1995	15	714.86	10.9	1.44	1.03
1996	21	376.11	12.62	1.65	1.26
1997	23	1021.02	6.83	1.43	1.3
1998	39	1054.24	8.89	1.59	1.5
1999	32	1053.6	7.59	1.51	1.38
2000	36	1591.74	16.05	1.95	1.69
2001	23	1124.85	23.24	1.84	1.5
2002	25	1114.7	20.11	1.88	1.44
2003	17	1616.8	12.95	1.71	1.38
2004	30	1191.08	9.4	1.67	1.35
2005	62	1667.45	10.62	1.61	1.22
2006	80	2919.73	6.83	1.57	1.11
2007	80	2454.18	9.66	1.59	0.99
2008	72	2501.49	13.58	1.78	1.04
2009	37	1353.34	15.76	1.81	1.13
2010	46	754.69	12.85	1.71	1.09
2011	59	1652.15	12.97	1.57	1.04
2012	70	1557.51	17.44	1.67	1.2
2013	74	1290.26	16.13	1.5	1.15
2014	71	1884.72	18.03	1.46	1.19
2015	70	2080.83	17.06	1.32	1.12
2016	23	2954.45	21.52	1.31	1.09

Table A.2: Venture Capital Funds

Vintage Year	Number of Funds	Average Size ($M)	Average Net IRR(%)	Average MOIC	Average PME
1980	0	-	-	-	-
1981	0	-	-	-	-
1982	1	54.8	8.4	1.74	0.72
1983	0	-	-	-	-
1984	2	52.5	9.27	1.78	0.89
1985	1	64.7	19.87	2.9	1.6
1986	1	42.7	8.2	1.38	0.79
1987	4	213.07	13.24	2.29	1.22
1988	0	-	-	-	-
1989	3	66	12.38	1.88	0.98
1990	7	111.44	17.07	2.15	1.16
1991	3	149.94	36.22	2.83	1.61
1992	8	114.76	30.49	3.8	1.9
1993	8	110.08	35.22	3.76	1.9
1994	9	119.44	36.9	6.21	3.01
1995	11	157.09	51.88	4.38	2.61
1996	16	183.05	37.27	3.36	2.42
1997	16	140.91	31.34	1.91	1.63
1998	27	250.46	20.91	1.69	1.6
1999	35	378.9	-2.48	0.81	0.78
2000	75	413.57	-3.24	0.94	0.77
2001	40	477.82	1.96	1.28	0.98
2002	24	279.22	-1.13	1.02	0.75
2003	18	252.99	0.35	1.15	0.82
2004	24	234.5	-0.78	1.49	1.06
2005	33	277.23	1.88	1.45	0.95
2006	44	466.26	2.41	1.35	0.83
2007	49	270.45	10.69	1.98	1.21
2008	40	398.29	8.12	1.85	0.99
2009	16	283.76	13.52	1.69	1.05
2010	18	310.71	13.21	1.78	1.13
2011	26	360.24	18.99	2.09	1.37
2012	18	429.33	12.26	1.65	1.16
2013	22	360.43	17.14	1.57	1.24
2014	26	313.71	31.34	1.64	1.27
2015	27	451.41	14.72	1.32	1.11
2016	7	191.56	22.05	1.43	1.11

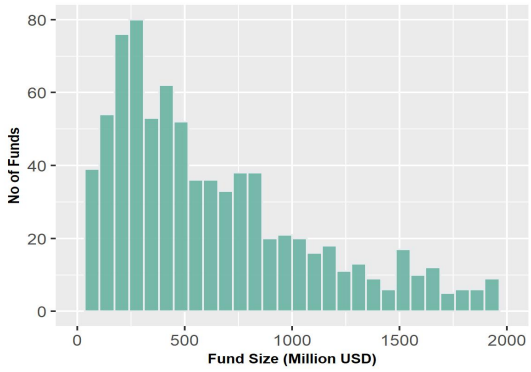
Note: Here we present summary statistic on fund size and performance for our sample of **1058** Buyout (BO) **659** Venture Capital (VC) funds segregated by vintage year which is defined as year when the fund made its first capital call from the LPs. The PME is constructed using the method presented by Kaplan and Schoar (2005). The data is sourced from Preqin's private equity data based accessed via WRDS. The last reported value for cash flows and NAVs was 30th June 2019.

A.2 List of Predictors

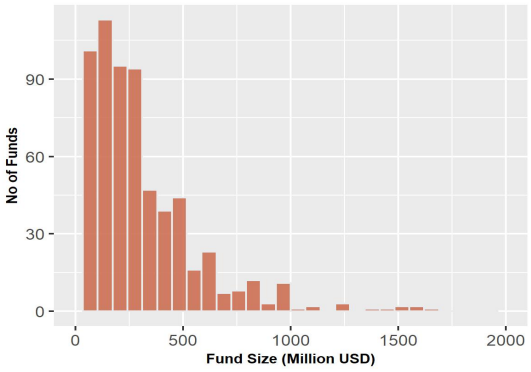
Variable	Type	Description	Data Source
PME	Binary	Takes the value 1, if the fund's KS-PME is greater than one, 0 otherwise	Preqin
Fund_Size_USD	Continuous	Targeted capital commited for the fund (in Million UDS)	Preqin
Fund_Number_Overall	Discrete	Total Number of funds raised by the fund manager/GP. Value of 1 indicates a first time fund	Preqin
Fund_Number_Series	Discrete	Total Number of funds raised in specific series by the fund manager/GP. Value of 1 indicates a first in series	Preqin
Firm_Age	Discrete	Difference between the vintage year of the fund and the year the GP firm was established	Preqin
Geographic_Scope_Diversified	Binary	Takes the value 1, if fund invests in more than one country, 0 if it is a country specific fund	Preqin
Industry_Diversified	Binary	Takes the value 1, if fund invests in multiple industries, 0 if it has a specific industry focus	Preqin
VC_Specialized	Binary	Used only for VC funds: Takes the value 1, if the fund speciliases in a particular financing stage, 0 otherwise	Preqin
GDP_yoy	Continuous	The nominal growth rate (YoY) of US GDP in the year preceding the fund's vintage year	Fred
TR_10yrs	Continuous	10-year US treasury bond yield during the year preceding the fund's vintage year	Fred
yoy_MSCI	Continuous	Annual returnof the MSCI world index for the year preceding the fund's vintage year	MSCI
Funds_Raised_Last_Year	Discrete	Number of funds raised in the year preceding the fund's vintage year	Preqin
Pcent_Increase_Funds_Last_Year	Continuous	Percentage Increase in the number of funds raised in the year preceding the fund's vintage year	Preqin
Fund_Focus_US	Binary	Takes the value 1, if fund primarily invests in the US, 0 otherwise	Preqin
Fund_Focus_Europe	Binary	Takes the value 1, if fund primarily invests in the Europe, 0 otherwise	Preqin
Fund_Focus_Asia	Binary	Takes the value 1, if fund primarily invests in the Asia, 0 otherwise	Preqin
Local_Currency_USD	Binary	Takes the value 1, if fund was raised in the US, 0 otherwise	Preqin
Local_Currency_EUR	Binary	Takes the value 1, if fund was raised in Europe, 0 otherwise	Preqin

Note: Here we present list of raw variables used to train our models, their numerical type and respective data sources. We also included squared terms for Fund_Size_USD, Fund_Number_Overall and Fund_Number_Series in our models to capture the potential concave/convex relationships observed between them and fund performance in past literature.

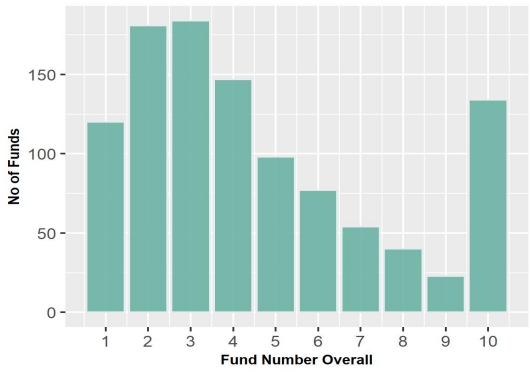
A.3 Statistics and Distributions: Predictor Variables



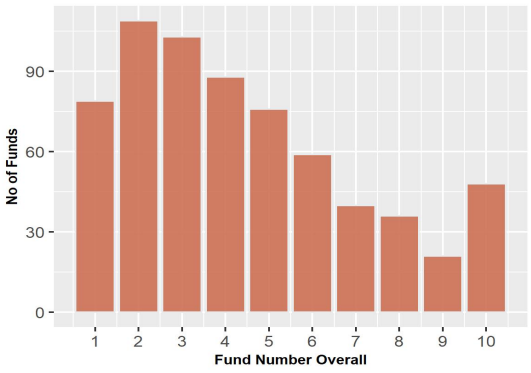
(a) Fund Size($M) - BO



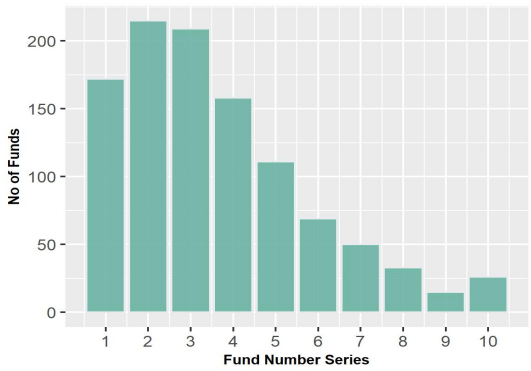
(b) Fund Size($M) - VC



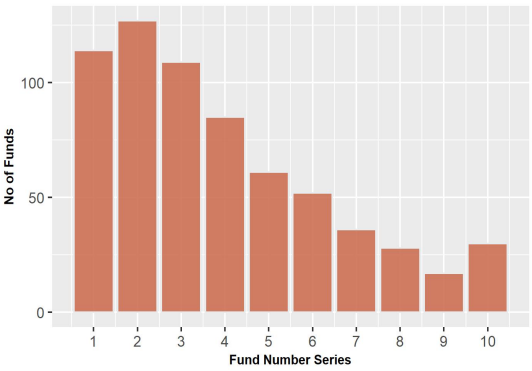
(c) Fund Number Overall - BO



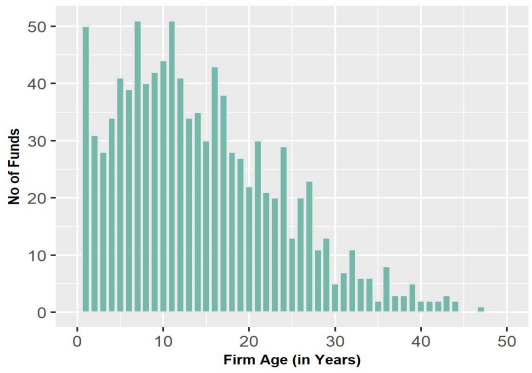
(d) Fund Number Overall - VC



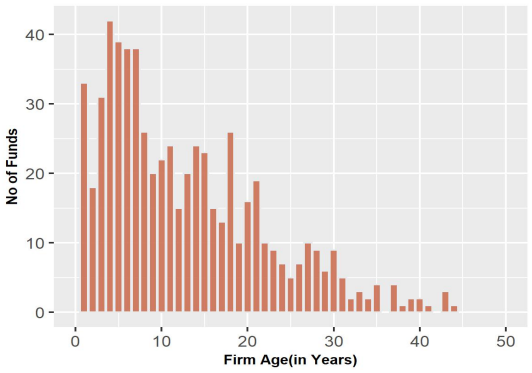
(e) Fund Number Series - BO



(f) Fund Number Series - VC

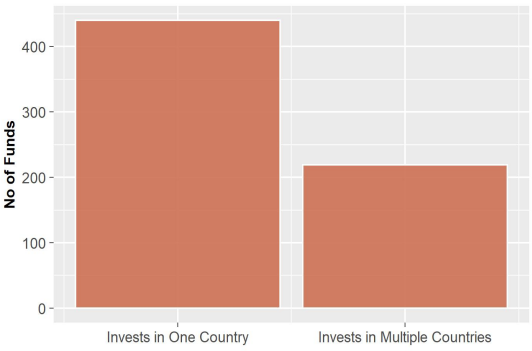


(a) Firm Age -BO

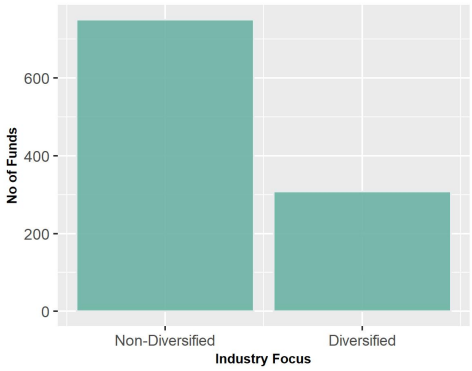


(b) Firm Age - VC

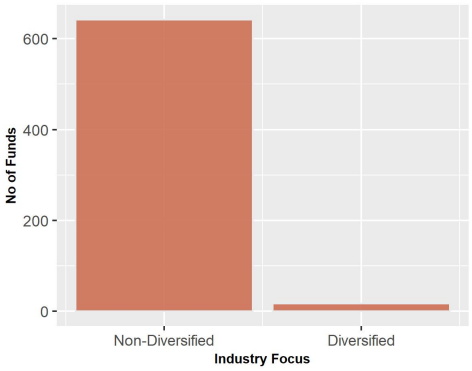
(c) Geographical Diversification -BO



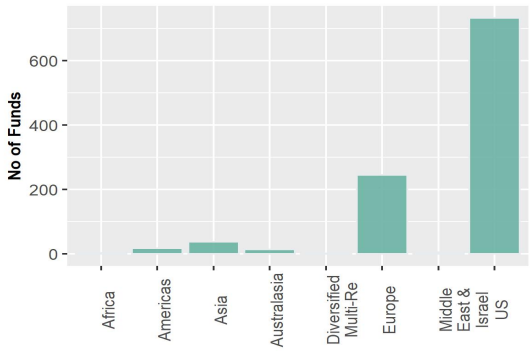
(d) Geographical Diversification - VC



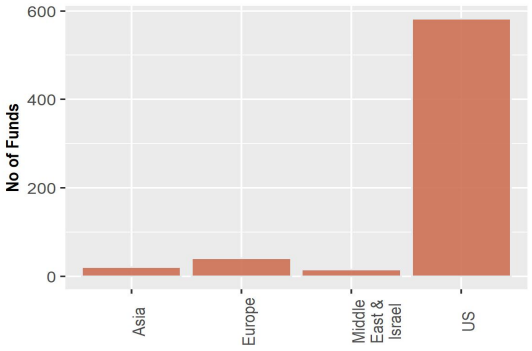
(e) Industry Diversification - BO



(f) Industry Diversification - VC



(g) Fund Focus - BO



(h) Fund Focus - VC

Table A.3: Predictor Statistics: BO Funds

	GDP_yoy	TR_10yrs	yoy_MSCI	Funds_Raised_Last_Year	Pcent_Increase_Funds_Last_Year
Minimum	-1.790	1.800	-42.080	3	-0.370
Maximum	11.730	12.460	39.110	2,266	0.910
Mean	4.640	4.160	8.740	956.310	0.180
Median	4.610	4.270	9.550	1,058	0.190
Stdev	1.910	1.600	15.700	481.140	0.230
Skewness	-1.400	0.720	-1.190	0.130	-0.150
Kurtosis	3.580	1	1.960	-0.210	0.230

Table A.4: Predictor Statistics: VC Funds

	GDP_yoy	TR_10yrs	yoy_MSCI	Funds_Raised_Last_Year	Pcent_Increase_Funds_Last_Year
Minimum	-1.790	1.800	-42.080	16	-0.370
Maximum	12.240	13.920	39.110	2,266	0.910
Mean	5.090	4.800	9.630	790.990	0.200
Median	5.660	4.800	12.840	764	0.220
Stdev	1.770	1.620	16.450	448.570	0.240
Skewness	-1.500	0.630	-0.970	0.500	-0.060
Kurtosis	4.590	2.290	0.790	0.050	0.230

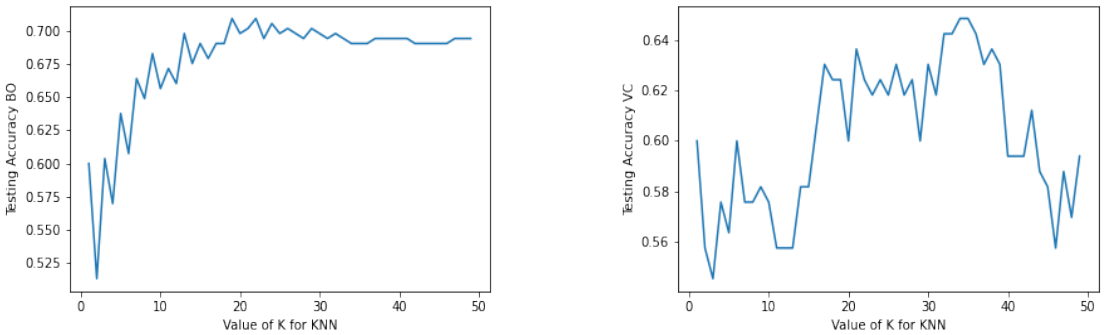
Appendix B

B.1 Hyper-Parameter Analysis

Here we discuss how change in hyper-parameters impacts the accuracy of our results.

Overall, no one method can dominate in every scenario and each method's applicability depends on the unknown decision function of the data-set. Generally speaking, if the true decision boundary is linear,then models like Linear Discriminant Analysis and Logistic Regression tend to perform well. When the boundaries are moderately non-linear, Quadratic Discriminant Analysis may give better results. And as the decision boundary gets complicated with additional non-linearity, non-parametric approaches such as KNN can perform well. For instance, we have used n = 5, the number for nearest neighbours for plotting the graphs. However, if we investigate the change the accuracy level by changing the number of the nearest neighbours from 1 to 50 then we see how in BO data-set the accuracy reaches a peak and then flattens out, whereas it drops after reaching the peak in the VC data-set. Such demonstration are typical cases in many data science problems.

We first employed the nearest neighbour algorithm to test how non-parametric techniques fare compared to our other models. KNN is a non-parametric method and therefore no assumptions are required to be made with respect to its decision boundary. Generally, KNN performs better than linear models when the decision boundary is non-linear. The case we do not observe in BO funds but is somewhat visible in VC funds.

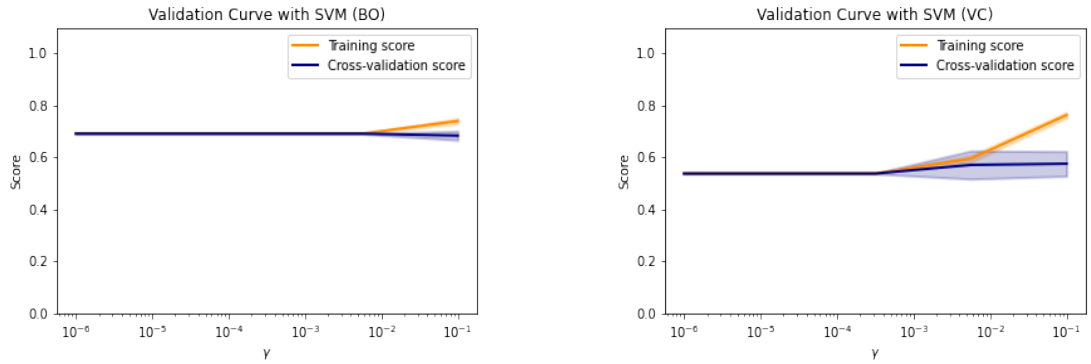


(a) Buyout (b) Venture Cap

Figure B.1: KNN Comparison

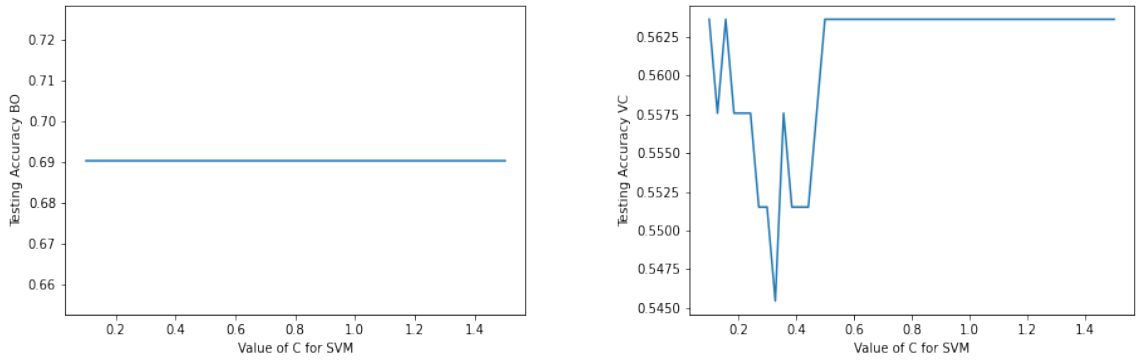
Coming to the analysis of the Support Vector Classifier model, we demonstrate in the plot B.2 the

training scores and validation scores of an SVM (non-linear) for different values of the kernel parameter gamma. We perform this analysis to re-establish that even non-linear kernels did not improve our accuracy for the model. Both the training score and the validation score are low, representing under-fitting. A moderate value of gamma will result in high values for both scores, i.e. the classifier is performing fairly well. However, if the gamma is too high, the classifier will over-fit resulting in a good training score but a poor validation score. We can only observe slight over-fitting in both the data-sets when the orange curve (Training score) is higher than the blue curve (Cross-validation score).



(a) Buyout (b) Venture Cap

Figure B.2: SVM Validation Curves



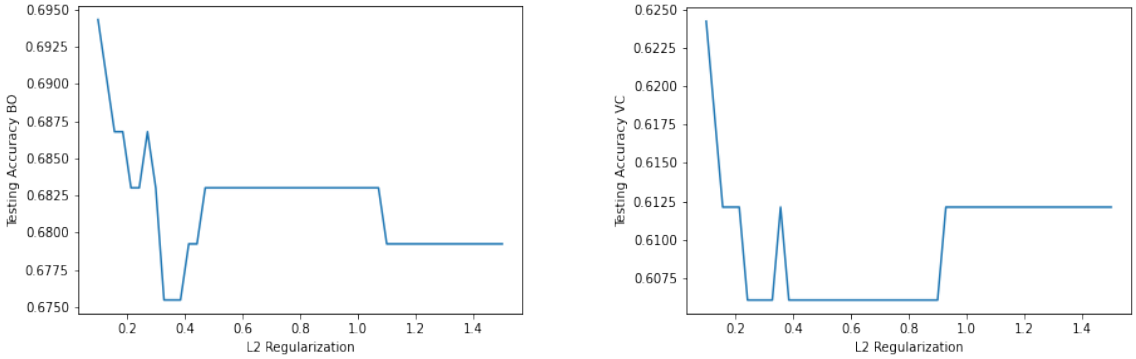
(a) Buyout (b) Venture Cap

Figure B.3: SVC Regularization Curves

Similarly we have also varied "C" i.e. the Regularization parameter to find the behaviour of SVC and Logistic models. Please note that the strength of the regularization is inversely proportional to C. In a sense, the "C" parameter indicates to the SVM optimization the degree to which misclassifying needs to be avoided in each training example. Hence, for large values of "C", the optimization will choose a smaller-margin hyperplane if it does a better job in getting all the training points classified correctly. Conversely, a very small value of "C" will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. For very tiny values of C, one usually

gets misclassified examples, often even if the training data is linearly separable. We observe that after a particular level of "C", the accuracy flattens out demonstrating the usage of regularization up to a particular value and checking for overfitting.

Similarly, the plots in B.4 demonstrates the change in model accuracy for Logistic Regression with respect to changing L2 regularization parameter.



(a) Buyout (b) Venture Cap

Figure B.4: Log. Regression Regularization Curves

The basic neural network model we employed also matches the performance of the Logistic and Linear Discriminant Analysis for VC, and could have performed better given a larger data-set. We use neural network model only to extent of comparison with our other parametric and non-parametric methods as discussed before, since the variance for these advanced models can be very high when dealing with small data-sets like ours.

B.2 Correlation Graphs

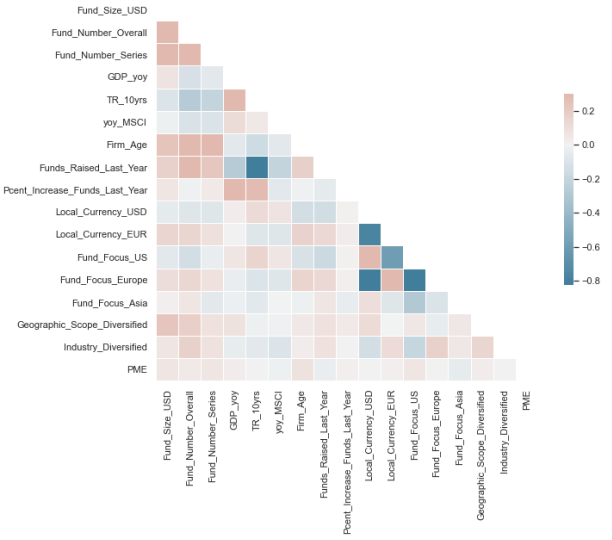


Figure B.5: Correlation graph - Buyout

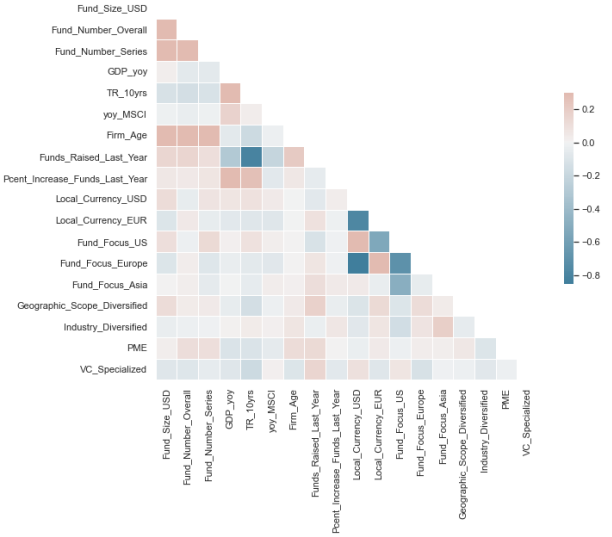
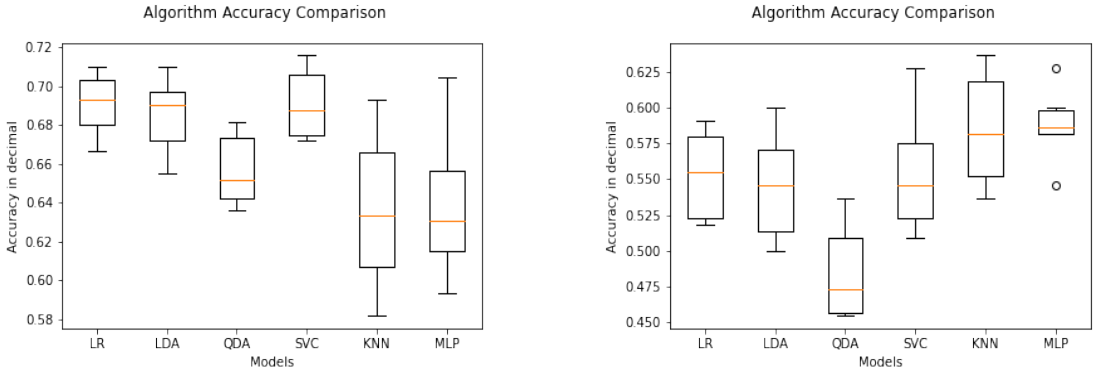


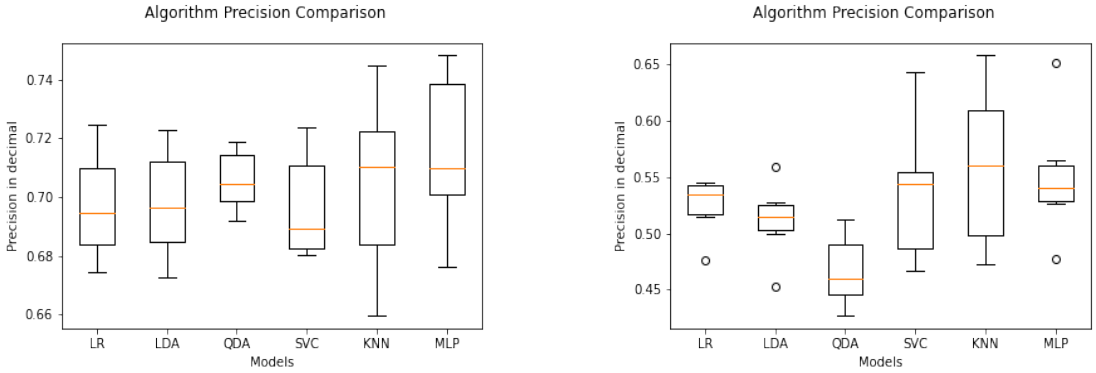
Figure B.6: Correlation graph - Venture Cap

B.3 Model Accuracy and Precision



(a) Buyout (b) Venture Cap

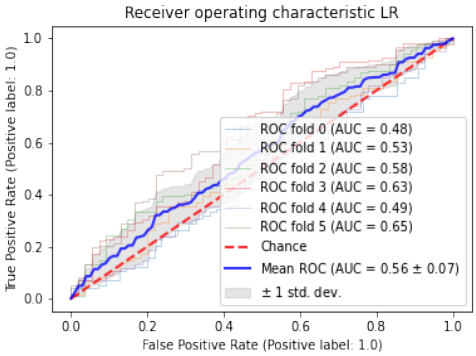
Figure B.7: Accuracy Comparison



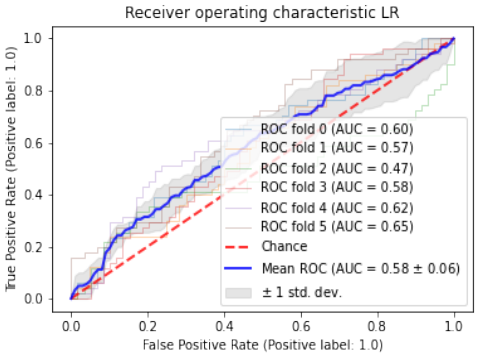
(a) Buyout (b) Venture Cap

Figure B.8: Precision Comparison

B.4 AUC-ROC Curves

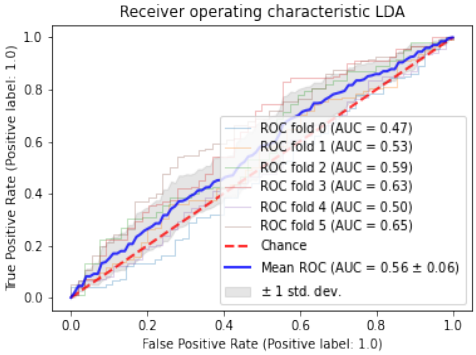


(a) LR Buyout

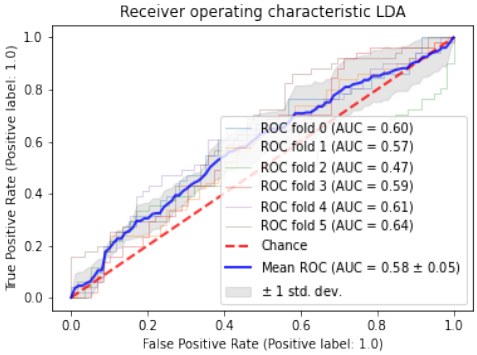


(b) LR VC

Figure B.9: LR Cross-Validation

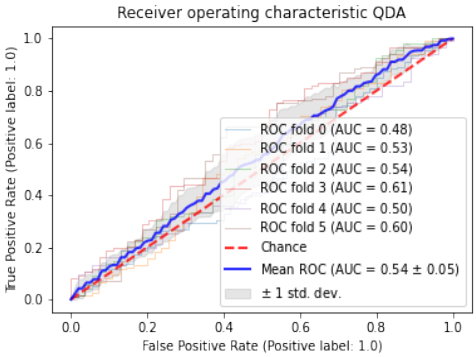


(a) LDA Buyout

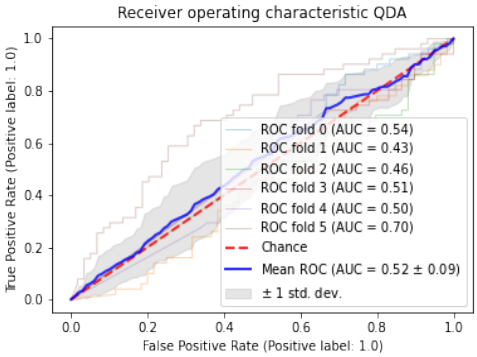


(b) LDA VC

Figure B.10: LDA Cross-Validation

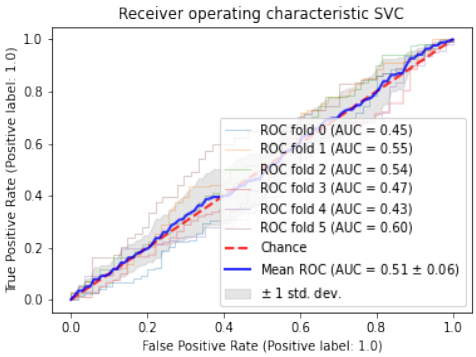


(a) QDA Buyout

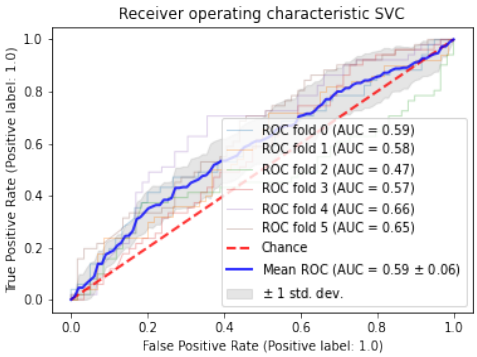


(b) QDA VC

Figure B.11: QDA Cross-Validation

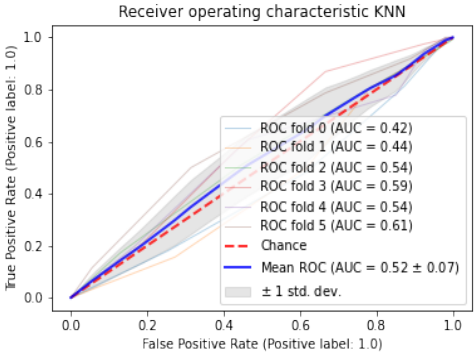


(a) SVC Buyout

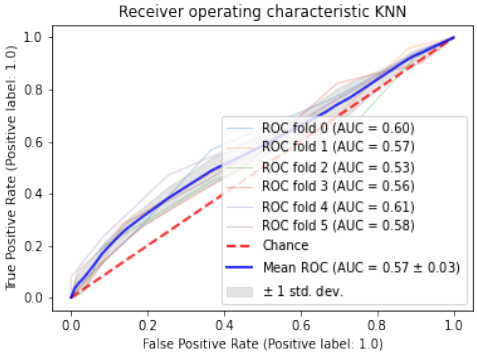


(b) SVC VC

Figure B.12: SVC Cross-Validation

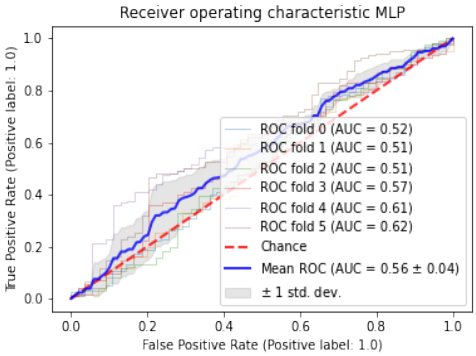


(a) KNN Buyout

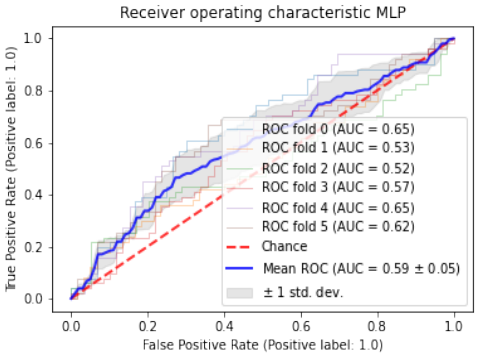


(b) KNN VC

Figure B.13: KNN Cross-Validation



(a) MLP Buyout



(b) MLP VC

Figure B.14: MLP Cross-Validation

Note: We have employed both cross-validation and stratification to find the mean AUC (Area under the Curve) for each model. We employed a 6-fold cross-validation for our analysis. Usually, the number for folds is higher if we have large data-sets. The grey area around the mean shows how each model runs and evaluates after each fold while gradually converging towards the mean.