COVERSHEET

THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

Faculty of Engineering, Computing and Mathematics

Assignment, Report & Laboratory Coversheet for Individual & Group Assignment

SUBMITTING STUDENT		
SURNAME SHELTON	**GIVEN NAMES** MARK ROBERT	**STUDENT NUMBER** 21151978
UNIT NAME HONOURS THESIS		**UNIT CODE** CITS4002
TITLE/TOPIC OF ASSIGNMENT DISSERTATION		**NAME OF LECTURER/TUTOR** TIM FRENCH
DATE/TIME DUE 29-MAY-2017	**DATE/TIME SUBMITTED** 29-MAY-2017	

HONOURS STUDENTS ONLY	OFFICE USE ONLY
By signing this document, I further assert that the length (word count) of my dissertation is within the maximum allowed length governed by the project unit I am enrolled in. Penalties, as outlined on this website, will be applied for over length dissertations.	

FOR GROUP ASSIGNMENTS ONLY NAME	STUDENT NUMBER
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

Unless other arrangements have been made it will be assumed that all group members have contributed equally to group assignments/laboratory reports

DECLARATION

I/We are aware of the University's policy on academic conduct (see over) and I/We declare that this assignment/project is my own/my group's work entirely and that suitable acknowledgement has been made for any sources of information used in preparing it. I/We have retained a hard copy for my/our own records.

SIGN:	**SIGN:**
SIGN:	**SIGN:**
SIGN:	**SIGN:**
SIGN:	**SIGN:**

NOTE: No assignment will be accepted without the declaration above being signed and dated

SEE OVER FOR INFORMATION ON REFERENCING & PLAGIARISM

REFERENCING

Information on appropriate referencing (citation) styles can be found under “Manage Your References” at: www.is.uwa.edu.au/information-resources/guides

Note that:
Each drawing, picture, photograph, quotation or block of text copied from a source must be acknowledged individually. This can be done using a referencing style (see above) or by including the full reference in the text or in a footnote. It is not sufficient to simply list sources in a bibliography at the end without including the individual references to the sources in the main text.

The same rules apply to materials taken from the web. The authorship and source must be traceable.
The boundaries between your original work and copied work must be clear. Use quotes, indentation and/or font style to make the distinction clear.

PLAGIARISM

“The appropriation or imitation of another’s ideas and manner of expressing them to be passed off as one’s own”.

(The Macquarie Dictionary, 1981)

Synonyms:

Piracy, copying, forgery, lifting, expropriation, appropriation

“Plagiarism is the unattributed use of someone else's words, creations, ideas and arguments as one's own. Within university policies it is usually further extended to include the use of 'too close' or extensive paraphrasing. For example, cutting and pasting text from the Web without attributing it to the author is plagiarism and therefore dealt with as cheating. Similarly, substituting a few words of copied text without changing the structure of the piece also constitutes plagiarism. There is a range of penalties for academic misconduct, depending on the seriousness of the cheating, from loss of credit to expulsion from the University.” (UWA Handbooks 2013)

The University of Western Australia treats plagiarism as serious academic misconduct. The University can impose severe penalties, including expulsion. Refer to Statue 17 Student Discipline and the associated Regulations for Student Conduct and Discipline at www.uwa.edu.au/current/information/discipline

See also

Faculty Policy on Plagiarism:

www.ecm.uwa.edu.au/students/exams/dishonesty

UWA’s policy statement on Ethical Scholarship, Academic Literacy and Academic Misconduct: www.handbooks.uwa.edu.au/postgraduate/policies

Towards Automated Venture Capital Screening

W.M.R. Shelton

This report is submitted as partial fulfilment
of the requirements for the Honours Programme of the
School of Computer Science and Software Engineering,
The University of Western Australia,
2017

Abstract

Venture Capital (VC) firms face the challenge of identifying a few outstanding investments from a sea of opportunities. The VC industry requires better systems to manage labour-intensive tasks like investment screening. Previous approaches to improve VC investment screening have common limitations: small, private datasets, a focus on early-stage investment, and narrow feature sets. To address these limitations, we present a multi-stage VC investment screening system. The system generates an optimised supervised learning classifier which it applies to data collected from large, public online databases (CrunchBase and PatentsView). We evaluate the system against three criteria: practicality, robustness, and versatility. The system satisfies each of these criteria. The system is practical in that it is near-autonomous. The system is robust in that it has only minimal variance in performance when trained on historical datasets. Finally, the system is versatile in that it addresses a large domain of investment prediction tasks with respect to forecast window, developmental stage and target outcome. This project also contributes a comprehensive empirical study of startup performance. The prior experiences of a startup's advisors, executives and founders are found to be the greatest predictors of startup performance. Ultimately, this project makes significant steps towards automation in the VC industry.

Keywords: Venture Capital, Investment Screening, Machine Learning
CR Categories: I.5, J.1, J.4

Acknowledgements

First and foremost I would like to thank my supervisors, Melinda Hodkiewicz and Tim French. Melinda and Tim are both fantastic academics and educators, and their strengths and talents perfectly complemented each other on this project. I am extremely grateful for Melinda and Tim's meticulous feedback and advice throughout the course of this year.

My time at university has been an incredible educational experience, and I primarily credit that to opportunities afforded to me by The University of Western Australia, the Fogarty Foundation, and St Catherine's College.

I would like to thank The University of Western Australia for guiding me through my, admittedly unusual, degree. Starting in Biomedical Sciences, I never imagined I would go on to complete a Double Major in Psychology and finally Honours in Computer Science. In particular, I would like to acknowledge the fantastic Bachelor of Philosophy (Honours) program, and Kathy Sanders and Jenna Mead for their advice and support.

I would like to thank the Fogarty Foundation. The Fogarty Foundation has provided support to me far beyond my UWA Fogarty Foundation Scholarship. The Fogarty Foundation was the catalyst that led to my involvement in Teach Learn Grow, encouraged me to start Bloom, and took me to Stanford University. In particular, I'd like to acknowledge Robyn King, Kathryn Clements, and Annie Fogarty for their incredible support throughout my degree.

I would like to thank St Catherine's College. In particular, Fiona Crowe and Mandy McFarland, who are truly inspiring leaders that have backed me and Bloom 110%. St Catherine's College is my extended family, and I am incredibly grateful for their support.

Finally, I would like to thank my friends and family who encouraged and supported me throughout my time at university.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Venture Capital (VC) is financial capital provided to early-stage, high-potential, high-growth companies (startups). VC firms have funded many successful companies, such as Google, Apple, Microsoft and Alibaba. Unlike investors in the public markets, VC firms often take a more active role in managing their investments, providing expertise and advice in both managerial and technical areas. VC firms have two primary roles: as scouts, identifying the potential of startups, and as coaches, helping startups realise that potential [1]. In these ways, the VC industry is critical to the success of startups and the commercialisation of new technologies more generally. Adoption of the Internet and inexpensive, ubiquitous computing has transformed the VC industry: companies require less funding to launch but more to scale in highly competitive markets [2]. There is now an impetus for VC firms to change the way they operate.

VC firms face the challenge of choosing a few outstanding investments from a sea of hundreds of thousands of potential opportunities. VC firms seek to make investments in companies that can provide a liquidity event that returns many times their investment value within the time-frame of their fund. For startups, a liquidity event (also referred to as an 'exit') is either an Initial Public Offering (IPO) or an acquisition by a larger competitor. Most VC firms expect their investments to exit within 3–8 years, per their fund time-frame [3]. When compared to public market investors, this is a long-term investment strategy. However, few companies are capable of maturing from early-stage to exit at this pace. In addition, traditional metrics of performance (e.g. cash-flow, earnings) often do not exist or are unclear [4]. VC firms must select from a field of many investment candidates, where little information is available on each of them, and only a few will grow at a fast enough rate to be worthwhile — this is why the VC investment screening process is considered difficult.

The VC industry is changing and requires better systems and processes to manage labour-intensive tasks like investment origination and screening efficiently. Currently, investment opportunities are either referred through a VC firm's networks or identified through technology scans (e.g. Google searches,

patent searches). These processes are time-consuming for VC firms. Attempts in the literature to solve this problem have three common limitations: small sample size [4, 5, 6, 7, 8, 9, 10], a focus on early-stage investment [11, 4, 12, 13, 10, 14], and a narrow feature set [4, 8, 12, 10, 9, 5]. Although individual studies address some of these limitations, none synthesise the findings into software ready for use in industry. The popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of the VC industry's desire for more quantitative methods of assessing startup potential [15]. There is preliminary evidence that mining these data sources may address previous limitations [14, 16].

We believe it is now possible to address previous limitations in this field and produce an improved VC investment screening system. Our system aims to identify startup companies that are likely to raise additional funding, become acquired or have an IPO (or some combination thereof) in a given period. This system could assist VC firms to efficiently screen investment candidates.

We assess our system against the following criteria:

1. Practicality. The system must be more efficient to use than manual investment screening. The system should be designed to operate with minimal user input and no assumed technical expertise. The system should also be designed to run in reasonable time.

2. Robustness. The system must be robust to changes over time. The system should be designed to have minimal variance in performance when training on datasets from different times so investors can trust its ability to make future-looking predictions. The system should also be designed to adapt to the quantum and type of data available from data sources over time.

3. Versatility. The system must be able to address a large domain of investment prediction tasks. The system should be designed to make accurate predictions for companies of different developmental stages (e.g. Seed, Series A), for different target outcomes (e.g. Acquisition, IPO) across different forecast windows (e.g. exit in two years, exit in four years).

We organise the thesis as follows:

- Chapter 2: Literature Review. We review the theoretical background of startup performance and VC investment and evaluate previous attempts to develop technologies for use in VC investment screening systems.

- Chapter 3: Design. We outline the methodology used to design our VC investment screening system.
- Chapter 4: Evaluation. We perform a series of experiments to evaluate our system against three criteria: practicality, robustness and versatility.
- Chapter 5: Discussion. We discuss the merits and limitations of our project and their implications for investors and future research.

CHAPTER 2

Literature Review

In this chapter, we first review the state of Venture Capital (VC) investment and develop criteria which we can use to evaluate VC investment screening systems. We then determine the best methodologies to adopt to develop such a system, which we break down into three areas: features, data sources and classification algorithms.

1. Business Context. VC firms perform screening on many potential investment candidates. Current screening methods are time-consuming and subject to human selection biases. Preliminary evidence suggests that VC firms would be likely to adopt an improved investment screening system that is practical, robust and versatile.

2. Features. VC investment is a key driver of startup development yet we have an incomplete understanding of the factors that influence VC investment decisions and the subsequent performance of those investments. Previous studies have explored a range of features, the most significant of which are human capital, economic conditions, and investment record. However, few individual studies have evaluated a comprehensive and diverse feature set.

3. Data Sources. Startup performance is a multi-faceted problem and different data sources provide different perspectives on its actors, relationships and attributes. We review online data sources which have the potential to provide large, diverse feature sets. Preliminary evidence suggests startup databases CrunchBase and AngelList are promising sources.

4. Classification Algorithms. Predicting startup performance is difficult. However, machine learning has been applied successfully in other areas of finance. We evaluate common classification algorithms with respect to their suitability for our problem and intended feature set. We conclude that Random Forests, Support Vector Machines and Artificial Neural Networks appear to be most suitable for VC investment screening systems.

2.1 Business Context

In this section, we provide an introduction to the business context of Venture Capital (VC) investment screening. We review VC firm strategy and their investment processes. We explore why the adoption of data mining in the VC industry has been slow and develop criteria by which we can evaluate whether a VC investment screening system is an improvement on current methods.

2.1.1 Venture Capital Industry

VC investment is a key enabler of technological innovation and critical to research and technology-intensive industries (e.g software, medical and agricultural technologies). VC investment is a form of private equity, a medium to long-term form of finance provided in return for an equity stake in potentially high-growth companies. VC investment, in comparison to other forms of finance, is characterised by a large number of investment candidates, high degree of uncertainty, a lack of reliable data on company performance (particularly financial performance), and extensive due-diligence.

VC firms are reliant on a small number of high-risk investments to produce outsized returns through successful exit events. A rule-of-thumb is that given a VC portfolio of ten investments: three will fail entirely, three will remain active but not be very profitable, three will be active and profitable, and one will be highly successful and provide the firm with a return on all of their investments [14]. Compared to other forms of finance, VC investment is biased towards control at the expense of risk mitigation. Although VC firms tend not to take majority stakes, they exert influence through significant minority stakes, board membership, and leveraging their networks [17].

2.1.2 Venture Capital Systems

The VC investment process involves several main stages: investment origination and screening, evaluation, structuring (e.g., valuation, term sheets), and post-investment activities (e.g., recruiting, financing). In this project, we are most interested in the first stage: origination and screening (which we refer to as 'screening'). This process is complicated. As the cost of starting businesses has decreased, investors are faced with an increasing number of investment opportunities to evaluate [2]. At the same, despite VC firms' considerable influence on the trajectory of their investments, they remain highly selective. Studies show VC investment rates vary between 1.5–3.5% of proposals considered [14].

Referrals from trusted sources (e.g. portfolio entrepreneurs, other investors) are often used by VC firm to initially screen opportunities. Some VC firms are known to ignore approaches that do not have a qualified referral. VC firms may also discover companies through papers, events and databases related to the startup scene or particular industries. These screening processes are labour-intensive and time-consuming, and even after a company is screened more due-diligence is needed before a final decision is made.

2.1.3 Evaluation Criteria

Despite evidence that VC firms might benefit from data mining, the VC industry has lagged behind other forms of finance (e.g. bond trading, loan applications, insurance) in adopting technology to aid their decision-making. Banks are able to evaluate loan requests in minutes while VC firms take far longer to put together deals, sometimes months. While these are markedly different forms of finance (VC has a longer return period, larger investments, higher risk profiles), a more data-informed and analytical approach to VC investment is foreseeable.

The VC industry's view on applying data mining techniques to investment screening can be broadly characterised as interested but cautious. VC firm adoption of databases like CrunchBase and AngelList has increased in recent years [15]. However, firms do not appear to be using these databases to build automated systems. Stone (2014) interviewed Fred Wilson of Union Square Ventures who said: "We have not been able to quantify [startup potential]. We haven't even tried. Although I am sure someone could do it and they might be very successful with it. To us, the ideal founding team is one supremely talented product oriented founder and one, two, or three strong developers, and nothing else." [14].

Based on our review of the VC industry and current VC screening processes, we have developed criteria by which we can evaluate VC investment screening systems:

1. Practicality. An improved VC investment screening system must be more practical than manual investment screening (e.g. referrals, research papers, Google search). If the system is not easy to use then it is unlikely to be adopted because of inertia, the cost of re-training staff, and limited technical expertise in the industry. Most VC firms are small in headcount so these effects may have more impact than in other forms of finance (e.g. banking). An ideal system might be one that runs in the background and makes recommendations on the front-end. This could strike a good balance between being helpful while not taking away control from VC firm [17].

Investment screening is not considered a time-sensitive process in the VC industry (unlike structuring, for example), but screening systems should be designed to process new data quickly enough that the predictions are up-to-date and relevant.

2. Robustness. An improved VC investment screening system must be robust to changes over time. VC firms have concerns over the quality and volatility of factors used in data mining to predict startup performance. Chris Dixon of Andreessen Horowitz stated: "I've seen a few attempts to do it quantitatively but I think those are often flawed because the quantitatively measurable things are either obvious, irrelevant, or suffer from over-fitting." [14]. Therefore, a screening system should have minimal variance in performance when training on datasets from different dates so investors can trust its ability to make future-looking predictions. In addition to robust performance, VC firms seek systems that are themselves robust to time and will not become quickly outdated. Systems should be able to adapt to new data sources and feature sets as they become available.

3. Versatility. An improved VC investment screening system must be able to address a large domain of investment prediction tasks. VC firms vary in the investments they make according to their interests, the life-cycles of their funds, and the portfolios that they hold [3]. For example, VC firms make investment decisions with different periods so they can strategically manage the investment horizons of their funds. An investment screening system should be versatile enough to to make accurate predictions for companies of different developmental stages (e.g. Seed, Series A), for different target outcomes (e.g. Acquisition, IPO) across different forecast windows (e.g. exit in two years, exit in four years).

2.2 Features

In this section, we provide a review of studies performed into Venture Capital (VC) investment decisions and startup performance. In Table 2.1 we evaluate factors that have been indicated to influence VC investment decisions and might be relevant to VC investment screening systems.

VC investment decisions are made in an environment of informational asymmetry [4]. Given this context, VC investment decisions can be broken down into two main components: the underlying determinants of startup performance (which are difficult to observe), and signals that correlate with startup perfor-

Features	Results from Studies	
	Significant	Non-Significant
Startup Potential		
Human Capital		
Founder Capabilities	[11, 8, 5]	[18, 19]
Advisor Capabilities	[1]	[4, 8]
Executive Capabilities	[11, 8, 19]	[4]
Social Capital		
Social Influence	[11, 8, 12, 7]	-
Strategic Alliances	[1]	-
Structural Capital		
Patent Filings	[6, 20, 1]	[4, 5]
Investment Confidence		
Third Party Validation		
Investment Record	[4, 11, 10, 6, 19]	-
Investor Reputation	[8, 9, 20]	[6]
Media Coverage	[11]	[8]
Historical Performance		
Financial Performance	[11, 1]	-
Non-Financial Performance	[8, 5]	[6]
Contextual Cues		
Industry Performance	[18, 10, 5]	[11, 19]
Broader Economy	[11, 10, 6, 19, 20]	[18, 4]
Local Economy	[18, 11, 10, 5, 6]	-

Table 2.1: Features relevant to VC investment investment. We review thirteen empirical studies that investigate drivers of VC investment. For each study, we note whether included features have a significant effect on the VC investment model.

mance (which are easier to observe). We will review factors that underpin these two components in the following sections.

2.2.1 Startup Potential

Determinants of startup performance from the literature can be broadly categorised into three areas: human capital, social capital and structural capital [1, 4].

- Human Capital. Human capital is critical to early-stage startups that have limited resources and are changing constantly. The education background of founders [11, 5], and the past entrepreneurial experience of advisors [1] and founders [5] have been linked to startup performance, though some studies dispute this [18, 19].

- Social Capital. Entrepreneurship is achieved through opportunity discovery and realisation, which occurs through the medium of social networks. Presence and engagement on Facebook and Twitter are predictive of startup performance [12, 11] and strategic alliances have also been found to predict VC investment [1].

- Structural Capital. Structural capital is the supportive intangible assets, infrastructure, and systems that enable a startup to function. Intellectual property and their proxy, patents, are a key component of structural capital for newly-formed startups. Patents and patent filings have been found to predict the survival and investment success of biotechnology startups [1, 6] but there is less supportive evidence for non-biotechnology startups [5, 4].

2.2.2 Investment Confidence

A key challenge of the VC investment process is informational asymmetry [4]. To get an understanding of the underlying potential of a company, investors may look to other factors to corroborate the evidence like third party validation, historical performance, and contextual cues.

- Third-Party Validation. Founders are optimistic about their startups so it is reasonable for investors to cross-reference their signals with third-parties. Third-party validation from credible sources like other investors [4, 11, 10, 6, 19], the media [11], and the government has been shown to factor into VC investors' decision-making processes.

- Historical Performance. Unlike in other forms of finance, it is challenging to measure the performance of VC candidates. Reporting is not standardised and profitability information is rarely available or too preliminary to be helpful. Simple performance metrics like survival time have been studied [8, 5] but are not helpful measures for VC investment screening.

- Contextual Cues. Startups do not exist in isolation but are rather a product of their context like any other business. Investors consider the performance of a startup's competitors [18, 10, 5], their local economy [11, 10, 5] and the broader economy [10, 6] when evaluating potential investment candidates.

2.2.3 Feature Evaluation

We collected evidence of features that influence startup performance and VC investment decisions. We found that previous studies typically focused on factors in isolation and have rarely evaluated a comprehensive feature set. Without a standardised evaluation methodology, this has led to considerable disagreement between studies as to which factors are important. We believe a diverse range of features is critical to developing accurate models of startup performance and investment decisions. We recommend that VC investment screening systems incorporate measures of the determinants of startup potential (human capital, social capital, and structural capital) and signals of investment confidence (third-party validation, historical performance, and contextual cues).

2.3 Data Sources

The identification of promising Venture Capital (VC) investment opportunities is a complex and difficult task. There are many factors that can influence VC investment decisions. Capturing the diversity of these factors is critical to developing accurate models. Appropriate selection of these data sources is important because different data sources provide insights into different actors, relationships and attributes. Ideally, tasks as complex as investment screening should involve data collection from multiple data sources.

Previous studies in this field have been limited by data sources restricted in sample size. Many studies have samples of fewer than 500 startups [4, 5] or between 500 and 2,000 startups [6, 7, 8, 9, 10]. Few studies have used larger samples (more than 100,000 startups), usually derived from CrunchBase or AngelList [18, 12]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [21]. Startup databases (e.g. CrunchBase) and social networks (e.g. Twitter) offer datasets larger than those used in many previous studies. We expect data collected from these sources will lead to the discovery of additional features and higher accuracy in VC investment prediction.

In Table 2.2, we outline the characteristics of relevant data sources and how they could contribute to features indicated to be relevant to VC investment decision-making. Furthermore, we describe desirable characteristics of data sources for VC investment screening, review potentially relevant data sources, and ultimately determine which data sources are most likely to suit the characteristics of VC investment screening.

Properties	Startup Databases		Social Media		Other Sources	
	CrunchBase	AngelList	LinkedIn	Twitter	PatentsView	PrivCo
Features						
Startup Potential						
Human Capital						
Founder Capabilities	✓	✓	✓✓	✗	✗	✗
Advisor Capabilities	✓	✓	✓✓	✗	✗	✗
Executive Capabilities	✓	✓	✓✓	✗	✗	✗
Social Capital						
Social Influence	✓	✓✓	✓✓	✓✓	✗	✗
Strategic Alliances	✓	✓	✗	✗	✓	✗
Structural Capital						
Patent Filings	✗	✗	✗	✗	✓✓	✗
Investment Confidence						
Third Party Validation						
Investment Record	✓✓	✓✓	✗	✗	✗	✓
Investor Reputation	✓	✓✓	✓	✗	✗	✗
Media Coverage	✓✓	✓	✗	✓	✗	✗
Historical Performance						
Financial Performance	✗	✗	✗	✗	✗	✓✓
Non-Financial Performance	✓✓	✓✓	✓	✗	✗	✓
Contextual Cues						
Industry Performance	✓	✓	✗	✗	✗	✗
Broader Economy	✓	✓	✗	✗	✗	✗
Local Economy	✓	✓	✗	✗	✗	✗
Ease of Use						
Cost Effective	✓	✓✓	✓	✗	✓✓	✗
Time Efficient	✓✓	✓✓	✗	✓✓	✓✓	✗
Accurate Data	✓	✓	✓✓	✓✓	✓✓	✓✓
Large Data Set	✓✓	✓✓	✓✓	✓✓	✓✓	✓

Table 2.2: Data sources relevant to VC investment screening. We reviewed six data sources commonly used in entrepreneurship research for their suitability for VC investment screening. We evaluated data sources on their ability to provide relevant features for our analyses and on their ease of use in data collection. We excluded offline sources from our analyses. Ratings are: ✗ = poor, ✓ = satisfactory, ✓✓ = good.

2.3.1 Source Characteristics

Entrepreneurship research is transforming with the availability of online data sources: databases, websites and social networks. Entrepreneurship studies have historically relied on surveys and interviews for data collection. Measures of human capital (e.g. founders' capabilities), strategic alliances, and financial performance are difficult to capture elsewhere. However, the trade-off for access to these features is that surveys and interviews are time-consuming and costly to implement. While online surveys address some of these issues, it is still difficult to motivate potential participants to contribute. Online data sources like startup databases and social networks are efficient because collecting data is a secondary function of users interacting with these sources. Researchers can also collect data from these sources automatically and at scale. For these reasons, we only consider online data sources for inclusion in this study, specifically crowd-sourced startup databases (e.g. CrunchBase, AngelList), social networks (e.g. Twitter, LinkedIn), government patent databases (e.g. PatentsView) and private company intelligence providers (e.g. PrivCo). We review the characteristics of each of these data sources commonly used in entrepreneurship research in Appendix A.

2.3.2 Source Evaluation

Entrepreneurship and VC investment research is primed to take advantage of new online data sources. We evaluated relevant data sources for their suitability to predicting startup investment. Startup databases CrunchBase and AngelList provide the most comprehensive set of features, including information on funding rounds, acquisitions, IPOs, employees, and investors. Both databases provide hundreds of thousands of company entries. There are small differences between the features recorded by each. CrunchBase has slightly more coverage but lacks AngelList's social network. At least one startup database should be used and either are satisfactory. Of the other data sources we reviewed, PatentsView is the most promising. PatentsView provides comprehensive patent information, though it could prove difficult matching identities to other sources. Other data sources are less promising because of access issues. LinkedIn cannot be easily collected now the API is deprecated. Twitter provides social network topology and basic profile information through its free API but does not provide access to historical tweets. Financial reports are too expensive for the purposes of this study.

2.4 Classification Algorithms

Predicting startup performance is a difficult problem for humans. However, in recent years, machine learning has been used successfully in other forms of finance and there may be scope to apply similar techniques to improve Venture Capital (VC) investment screening. Machine learning is characterised by algorithms that improve their ability to reason about a given phenomenon given greater observation and/or interaction with said phenomenon. Mitchell (1997) provides a formal definition of machine learning in operational terms: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." [22].

Machine learning algorithms can be classified based on the nature of the feedback available to them: supervised learning, where the algorithm is given example inputs and desired outputs; unsupervised learning, where no labels are provided and the algorithm must find structure in its input; and reinforcement learning, where the algorithm interacts with a dynamic environment to perform a certain goal. These algorithms can be further categorised by desired output: classification, supervised learning that divides inputs into two or more classes; regression, supervised learning that maps inputs to a continuous output space; and clustering, unsupervised learning that divides inputs into two or more classes.

We evaluated common machine learning algorithms with respect to their suitability for use in identifying startup investment potential. In Table 2.3, we rank these algorithms by cross-referencing their assumptions and properties with the task characteristics. In the following sections, we describe the characteristics of VC investment screening, review common machine learning algorithms, and determine which algorithms are most likely to suit the characteristics of VC investment screening.

2.4.1 Task Characteristics

Machine learning tasks are diverse. VC investment screening is a task that suits supervised machine learning algorithms. We have access to historical labelled data, in the sense that we can use measures like whether a company has been acquired as a label of success. The key objective of machine learning algorithm selection is to find algorithms that make assumptions consistent with the structure of the problem (e.g. tolerance to missing values, mixed feature types, imbalanced classes) and suit the constraints of the desired solution (e.g. time available, incremental learning, interpretability). In the following sections, we outline the

Criteria	Machine Learning Algorithms													
	NB		LR		KNN		DT		RF		SVM		ANN	
Data Set Properties	**2**		4		6		**2**		**1**		4		6	
Missing Values	✓✓	[23]	✓	-	✗	[23]	✓✓	[23]	✓✓	[24]	✓	[23]	✗	[23]
Irrelevant Features	✗	[23]	✗	[25]	✓	[23]	✓✓	[23]	✓✓	[24]	✗	[23]	✗	[23]
Imbalanced Classes	✓✓	-	✓✓	-	✗	-	✗	[23]	✓	[24]	✓✓	[23]	✓	[23]
Algorithm Properties	**2**		**1**		4		4		**2**		6		6	
Predictive Power	✗	[21]	✓	[21]	✓	[21]	✗	[23]	✓✓	[21]	✓✓	[21]	✗✓	[21]
Interpretability	✓✓	[23]	✓✓	[25]	✗	[23]	✓✓	[23]	✓	[25]	✗	[23]	✗	[23]
Processing Speed	✓✓	[23]	✓✓	[21]	✓✓	[23]	✓	[23]	✓	[21]	✗	[23]	✗	[23]
Overall	**2**		**2**		6		4		**1**		5		7	

Table 2.3: Evaluation of machine learning algorithms for use in VC investment screening. We reviewed seven common supervised machine learning algorithms for their suitability in VC investment screening. We evaluated algorithms for their robustness to the structure of the dataset and their appropriateness for the constraints of our implementation. We ranked the algorithms according to the sum of these measures (in each section and overall) and emphasised highly-ranked algorithms. Ratings are: ✗ = poor, ✓ = satisfactory, ✓✓ = good. Algorithms are: NB = Naive Bayes, LR = Logistic Regression, KNN = K-Nearest Neighbours, DT = Decision Trees, RF = Random Forests, SVM = Support Vector Machines, ANN = Artificial Neural Networks.

characteristics of supervised learning tasks relevant to VC investment screening.

2.4.1.1 Data Set Properties

While datasets can be pre-processed to assist with their standardisation, some types of datasets are still better addressed by particular algorithms. Data set properties like missing data, irrelevant features, and imbalanced classes all have an effect on classification algorithms.

- Missing Values. Data sets often have missing values, where no data is stored for a feature of an observation. Missing data can occur because of non-response or because of errors in data collection or processing. Missing data has different effects depending on its distribution through the dataset. Public datasets, like startup databases and social networks, are typically sparse with missing entries despite their scale. Therefore, robustness to missing values is a desirable property of our algorithm.
- Irrelevant Features. Despite efforts to only include features that have theoretical relevance, machine learning tasks often include irrelevant features. Irrelevant features have no underlying relationship with classification. Depending on how they are handled they may affect classification or slow the algorithm. We expect irrelevant and non-orthogonal features in datasets used in VC investment screening because the features that predict startup performance have not been thoroughly evaluated in the literature. Therefore, robustness to irrelevant features is a desirable property of our algorithm.
- Imbalanced Classes. Data sets are not usually restricted to containing equal proportions of different classes. Significantly imbalanced classes are problematic for some classifiers. In the worst case, a learning algorithm could simply classify every example as the majority class. Startup datasets are likely to be highly imbalanced because very few startups are successful. Therefore, robustness to imbalanced classes is a desirable property of our algorithm.

2.4.1.2 Algorithm Properties

The desired properties of machine learning algorithms are related to the business problems that are being addressed. Predictive power, interpretability and processing speed are all desirable characteristics but involve trade-offs and must be prioritised.

- Predictive Power. Predictive power is the ability of a machine learning algorithm to correctly classify new observations. If a model has no predictive power, the model is not representing the underlying process being studied. For this reason, predictive power is a desirable property of our algorithm. However, if multiple algorithms provide similar predictive power other selection criteria become significant.

- Interpretability. Interpretability is the extent to which the reasoning of a model can be communicated to the end-user. There is a trade-off between model complexity and interpretability. Some models are a "black box" in the sense that data comes in and out but the model cannot be interpreted. For the purposes of investment screening, it is critical that VC firms understand the logic being used by the system so they can trust its predictions. Therefore, interpretability is a desirable property.

- Processing Speed. Finally, processing speed is another desirable property, especially when handling real-time data or when there is a need to run exploratory analyses on the fly. Generally VC investment decisions are made over weeks and months, but there is still a need for the system to quickly process new information as it becomes available.

2.4.2 Algorithm Characteristics

Supervised machine learning are algorithms that reason about observations to produce general hypotheses that can be used to make predictions about future observations. Supervised machine learning algorithms are diverse, from symbolic (Decision Trees, Random Forests) to statistical (Logistic Regression, Naive Bayes, Support Vector Machines), instance-based (K-Nearest Neighbours), and perceptron-based (Artificial Neural Networks). In Appendix B, we describe each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in applications relevant to VC investment.

2.4.3 Algorithm Evaluation

We evaluated supervised learning algorithms for their suitability for use in VC investment screening systems. While our evaluation gives directionality of fit, we hesitate to discard algorithms based on our literature review. Algorithm selection is complex and preliminary testing in the following chapter will provide clarity as to which algorithms should be used. In addition, larger training sets and good feature design tend to outweigh algorithm selection [21]. With those concessions

aside, our review suggests we should expect Random Forests, Support Vector Machines and Artificial Neural Networks to produce highest classification accuracies. An ensemble of these algorithms may improve accuracy further, though at the cost of computational speed and interpretability. We may expect Random Forests to outperform the other two algorithms because of robustness to missing values and irrelevant features and native handling of discrete and categorical data. However, Random Forests are not highly interpretable so Decision Trees and Logistic Regression may be preferable for exploratory analysis of the dataset.

2.5 Research Gap

As the cost of starting businesses decreases, Venture Capital (VC) firms are faced with an overwhelming number of investment candidates to assess and evaluate. The VC industry requires better systems and processes to efficiently manage labour-intensive tasks like investment screening. Attempts to address this problem have three common limitations: small sample size [4, 5, 6, 7, 8, 9, 10], a focus on early-stage investment [11, 4, 12, 13, 10, 14], and a basic feature set [4, 8, 12, 10, 9, 5]. In addition, there is little evidence that previous research has been translated into systems that are able to assist investors directly. We conducted a literature review to determine whether there was scope to address these limitations and produce a system that assists VC firms in screening investment candidates.

First, we reviewed the business context and developed criteria to guide the development and evaluation of our system: practicality, robustness and versatility. Second, we reviewed studies that have developed models of startup potential and realised few individual studies have evaluated a comprehensive and diverse feature set. Third, we assessed potential data sources and found preliminary evidence that suggests that the startup databases CrunchBase and AngelList are promising. Finally, we reviewed supervised machine learning techniques as applied to startup investment and other areas of finance. Our analyses suggested that we should expect Random Forests, Support Vector Machines and Artificial Neural Networks to be most suitable for our system.

This literature review provided evidence to suggest that it is possible to address previous limitations in this domain and produce an improved VC investment screening system that is practical, robust, and versatile. In the next chapter, we outline the process by which we developed that system.

CHAPTER 3

Design

In this chapter, we describe the methodology used to design our Venture Capital (VC) investment screening system. Figure 3.1 depicts the architecture of our system, structured into five components: data collection, dataset preparation, pipeline creation, pipeline selection, and prediction. We evaluate the performance of this system in the next chapter.

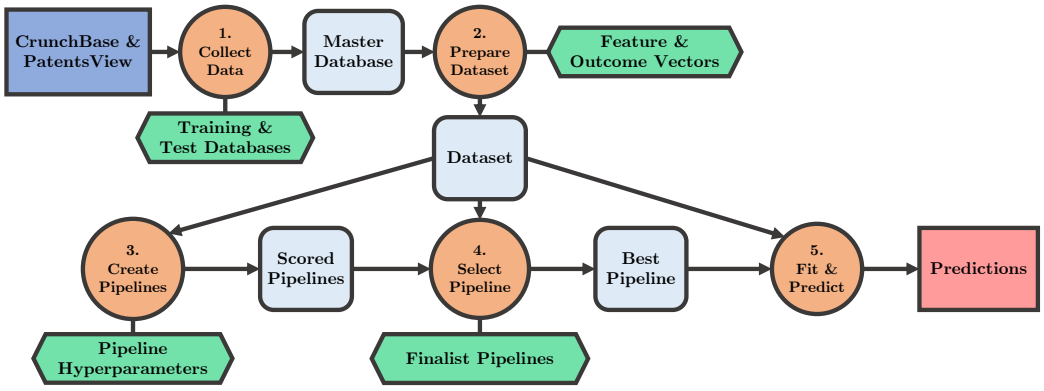


Figure 3.1: An overview of the system architecture proposed by this project. Legend: dark blue square = input, orange circle = system component, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

1. Data Collection. We developed a conceptual framework to guide our feature and data source selection. Based on this framework, we decided to collect data from CrunchBase and PatentsView. Our system collects data from these sources and imports it into a relational database.

2. Dataset Preparation. The system extracts database slices and converts them into datasets suitable for supervised learning. The system cleans the databases to ensure they only include relevant companies. We report the descriptive statistics of an indicative dataset.

3. Pipeline Creation. We identified issues of sparsity, long-tailed distributions, imbalanced feature ranges, and non-orthogonality in our datasets. We developed a classification pipeline system to address these dataset issues. Our system performs a grid search across the hyperparameters of the pipeline to generate scored, candidate pipelines.

4. Pipeline Selection. Next, our system selects the best candidate pipelines and evaluates their robustness over time. The outcome of this process is a single pipeline optimised to suit the dataset and prediction task. To evaluate robustness over time, we recreated historical datasets using a technique that filters records by their time-stamps.

5. Model Fit and Prediction. Finally, our system applies the optimised pipeline to a feature vector from a held-out test database, which generates a model and a set of predictions. We evaluate the models and predictions produced by our system in the next chapter.

3.1 Data Collection

In the previous chapter, we reviewed features and data sources used in entrepreneurship and Venture Capital (VC) investment research. In this section, we first discuss how we developed a conceptual framework to guide our feature and data source selection and then describe the process of collecting and storing data from CrunchBase and PatentsView.

3.1.1 Conceptual Framework

While previous studies into startup performance have explored a range of features, few individual studies have evaluated a comprehensive and diverse feature set. We developed a conceptual framework to ensure we included a comprehensive set of features in our VC investment screening system.

Our conceptual framework built on previous work by Ahlers (2015) to model investment decisions on equity crowd-funding platforms [4]. We sought to generalise Ahlers' framework beyond equity crowd-funding. While the first factor of their framework (venture quality) applies to startups generally, they defined their second factor (investment confidence) with respect to features specific to equity crowd-funding. We developed this second factor further, describing investment confidence as a product of third party validation, historical performance and contextual cues.

Our proposed conceptual framework is depicted in Figure 3.2. In the previous chapter, we described the features that underpin each factor and outlined the theoretical and empirical evidence that justify their inclusion in our conceptual framework.

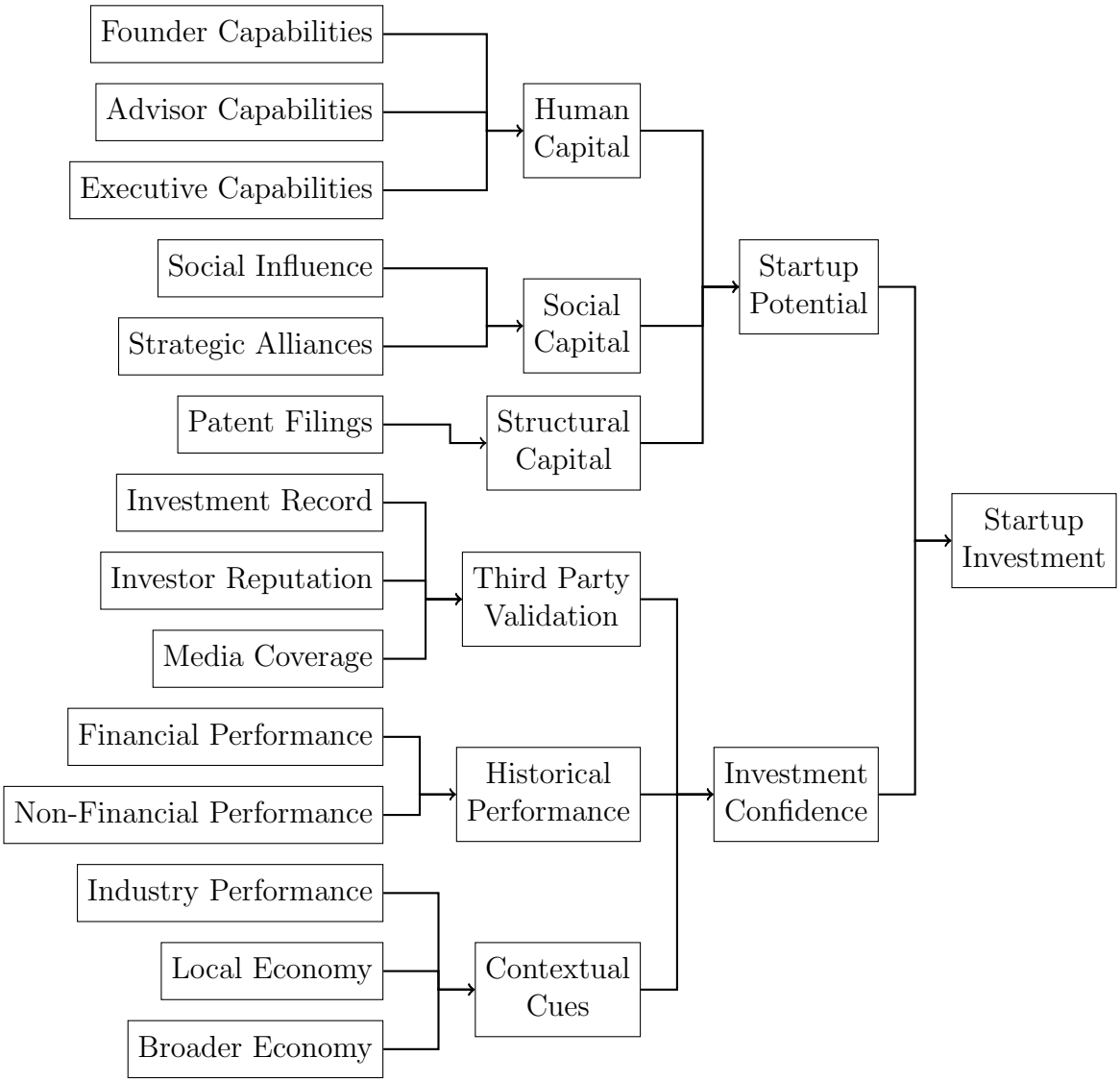


Figure 3.2: Proposed conceptual framework for VC investment. We adapted the framework proposed by Ahlers et al. [4], originally based on work by Baum and Silverman [1]. This framework includes features identified by the literature review in Chapter 2.

3.1.2 Data Sources

We sought data sources that could provide features that support the factors in our conceptual framework. We decided to collect data from CrunchBase with supplementation from PatentsView, a patent filings database. These sources cover the majority of factors in our conceptual framework. We discuss the process of collecting data from these sources, as depicted in Figure 3.3.

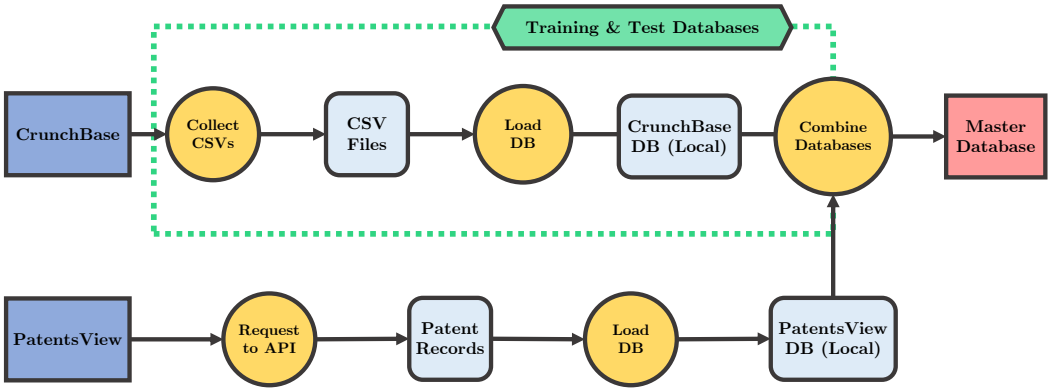


Figure 3.3: Data collection overview. CrunchBase data is collected multiple times to produce separate training and testing databases. Legend: dark blue square = input, yellow circle = process, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

3.1.2.1 CrunchBase

We were granted an Academic License to collect data from CrunchBase. CrunchBase provides database access in a few formats that offer trade-offs in terms of accessibility and comprehensiveness: REST API, CSV Dumps, MS Excel Summary. We chose to use the CSV Dumps because they provided a good trade-off between ease of use and comprehensiveness of access. CrunchBase provides a CSV file for each of CrunchBase's primary endpoints (e.g. organisations, people, funding rounds) which can be loaded easily into relational databases (see Appendix C for the database schema). We downloaded CSV Dumps from CrunchBase on 09 September 2016 and 04 April 2017 which became our training and testing databases, respectively.

3.1.2.2 PatentsView

We used PatentsView to obtain the patent filing records of each company in our CrunchBase dataset, focusing on information relating to dates, citations, and patent types. Our system matches companies across CrunchBase and PatentsView by standardising the company names (removing common suffixes, punctuation, etc.) and using normalised Levenshtein distances to determine similarity. Although approximate matching introduces error, the volume of companies in the database is too high to be matched manually, and there are no other identifying records. We stored the PatentsView data in a relation which we merged into our CrunchBase data to form our master databases.

3.2 Dataset Preparation

Our system performs multiple steps to prepare datasets from our training and test databases for use in machine learning, as depicted in Figure 3.4. In this section, we describe the process of preparing our datasets, which involves two components: generating historical databases from our relational database and converting these relational database slices into clean datasets ready for machine learning. Finally, we present the descriptive statistics of our dataset.

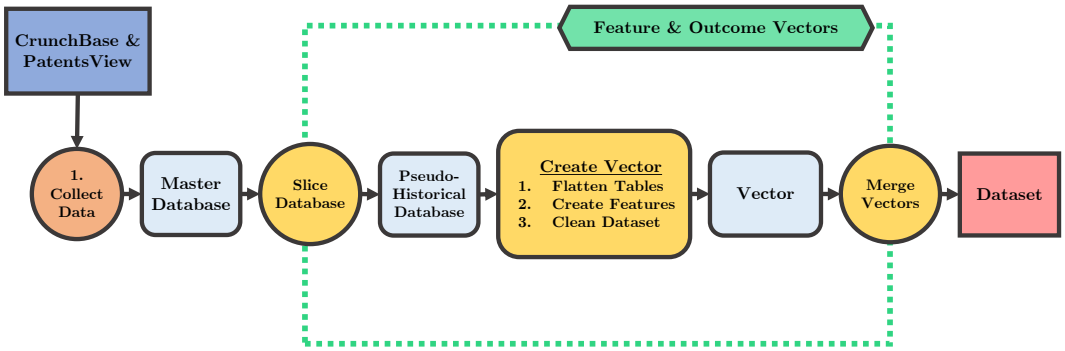


Figure 3.4: Dataset preparation overview. Feature and outcome vectors are created from the master relational database. Legend: dark blue square = input, yellow circle = process, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

3.2.1 Database Slicing

We developed a procedure for generating historical databases from our CrunchBase and PatentsView data. CrunchBase provides 'created-at' and 'last-updated' time-stamps for each record in their CSV-formatted dumps (and also in the JSON-formatted responses from their API). We used this to reverse-engineer previous database states by filtering the master database to only include records created before a given 'slice' date.

We evaluated our slicing technique by comparing a CrunchBase database collected in December 2013 with a slice engineered from our training database (collected in September 2016), as shown in Figure 3.5. There are only minor differences in the counts between the methods.

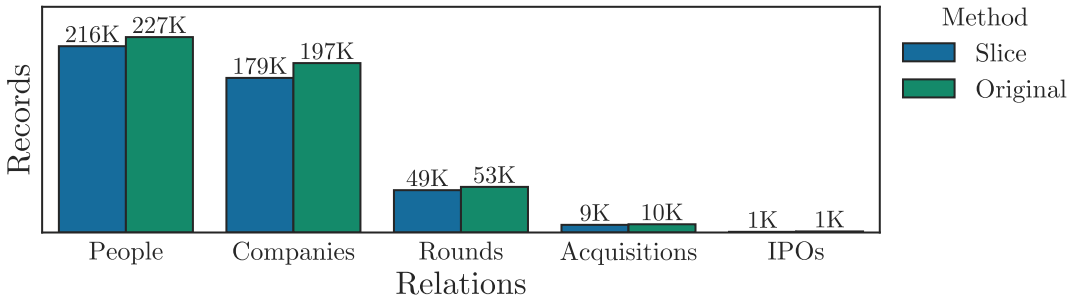


Figure 3.5: Database slice compared with original database. Original database collected in December 2013. Database slice generated from the training database collected in September 2016 and sliced to only include records created prior to December 2013.

Figure 3.6 presents company counts by startup development stage from different dataset slices. Dataset counts across all developmental stages have steadily increased over time. Before 2012 the datasets become too small to use to make meaningful predictions.

3.2.2 Vector Creation

To prepare our datasets for machine learning, our system performs aggregation, feature creation and preliminary screening. In the following section, we evaluate the effect of these processes on an indicative database sliced from the training database as of 09 September 2016 (N = 425,934).

First, our system flattens the relational database slices into a single file using Structured Query Language (SQL) aggregation queries. The system aggregates

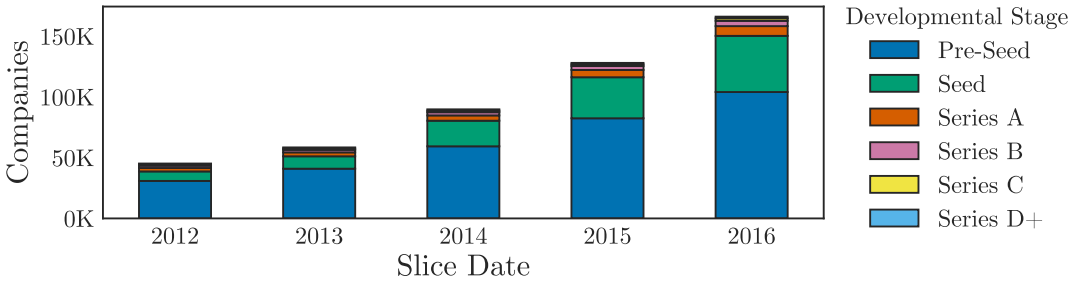


Figure 3.6: Dataset slice counts over time. Each slice was generated on April 04 of each respective year. Proportions by developmental stage stay relatively constant over this time-frame.

each relation in turn, grouping by Company ID and combining each aggregated table using left outer joins. Next, the system converts tuples (e.g. Round Type and Round Date) and lists (e.g. Round Types) into dummy variables.

Our system performs preliminary screening to remove traditional, non-startup businesses from the dataset (e.g. consulting firms, companies that will not take Venture Capital (VC) funding, etc.). To do this, we explored two factors for each company: developmental stage and age. By developmental stage, we primarily refer to external funding milestones. These stages also indicate shifts in a startup's functions and objectives. Our dataset as grouped by startup developmental stage is depicted in Figure 3.7.

After attempting to place the companies into development stages, a large group of companies that have not raised funding remain. We classify these companies into two groups – those that intend to raise funding and those that do not. We applied a cut-off equal to the 90th percentile of the age of companies in the Seed category and excluded the older group from further analyses (N = 227,162, 53.3%). As we are only interested in companies that could seek investment, we also excluded Closed, Acquired and IPO groups from further analyses (N = 35,973, 8.4%).

Figure 3.8 depicts the ages of companies in the dataset grouped by developmental stage. There is a positive relationship between age and developmental stage. Most pre-Series A companies are under five years old, and the majority of Series D+ funded companies are under ten years old, and the 75th percentile is at 15 years old. On this basis, we excluded companies that are over the 75th percentile for the age of companies in the Series D+ category (N =9,756, 2.2%). Our preliminary screening reduced the dataset from 425,934 companies to 153,043 companies, a reduction of 64.1%.

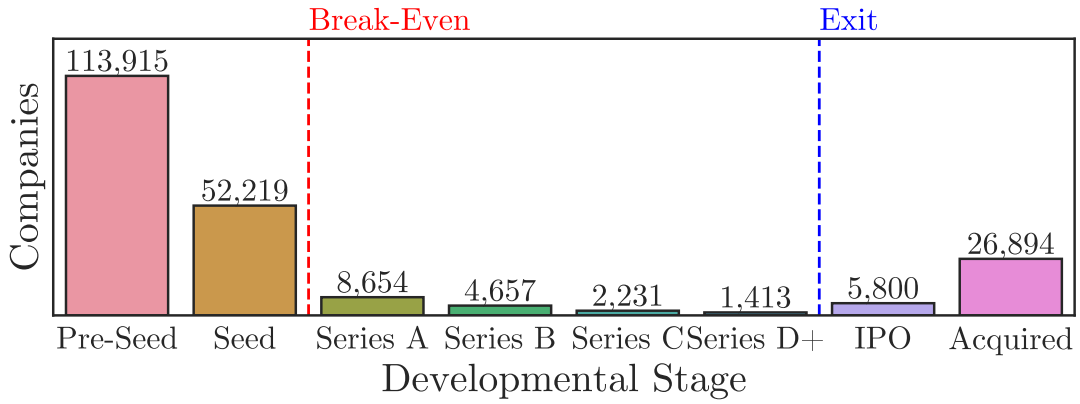


Figure 3.7: Companies grouped by stages of the startup development life-cycle. Companies at Pre-Seed and Seed stages are typically unprofitable and seek external funding to sustain their operations. Companies at Series A - Series D+ are typically either profitable or at least revenue-generating, and seek external funding to expand their operations.

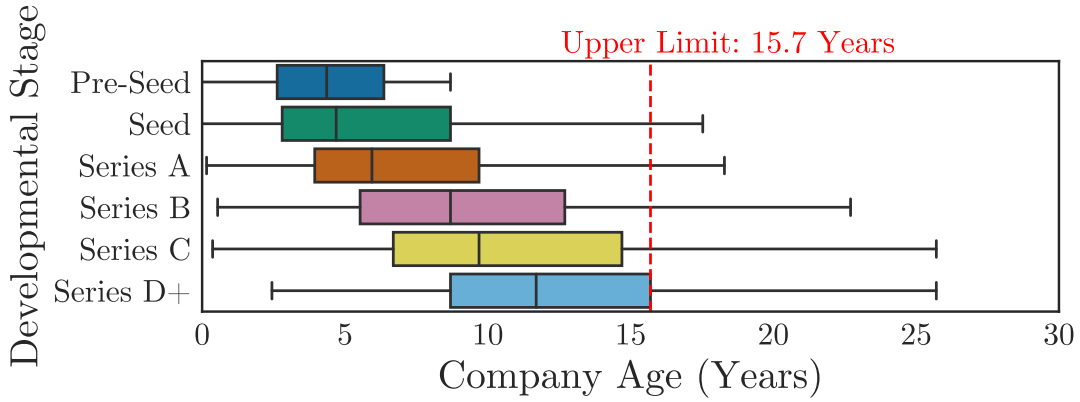


Figure 3.8: Company ages in years grouped by developmental stage. The dashed red line represents the 75th percentile of the age of companies in the Series D+ category (15.7 years).

3.2.3 Descriptive Statistics

Table 3.1 presents the descriptive statistics for the cleaned dataset from the previous section. The dataset skews towards Pre-Seed companies (i.e. companies that were recently founded and had not raised funding yet, 68.9%). These companies have few available features in comparison to later developmental stages. We investigate the impact of this on our predictions in Chapter 4. The interquar-

tile ranges imply significant variability in all measures. We do not believe that this indicates that the data has not been cleaned adequately, but rather, that it reflects that startup companies vary in their traits.

Stage	Obs	Age (Years)		Funding Raised (USD, M)		Funding Rounds (N)		Patent Filings (N)		Available Features (N)	
	N	50th	75th	50th	75th	50th	75th	75th	90th	50th	75th
Pre-Seed	113,915	4.36	6.36	0.00	0.00	0	0	0	0	25	133
Seed	38,942	4.66	6.69	0.25	1.30	1	2	0	1	178	231
Series A	6,615	5.69	8.70	4.40	9.41	2	3	0	2	239	302
Series B	3,342	7.61	10.70	14.89	28.20	3	4	0	4	255	314
Series C	1,610	8.70	11.70	35.29	62.00	3	5	1	9	305	321
Series D+	998	9.70	12.70	74.39	130.8	5	7	4	19	319	330
Included	165,422	4.69	6.69	0.00	4.00	1	2	0	1	90	160

Table 3.1: Descriptive statistics grouped by developmental stage.

CrunchBase's industry classification is simplistic compared to other databases (e.g. USSIC, VentureSource) which take a structured, hierarchical approach. For example, "Software", "Internet Services" could describe the majority of companies included in the database and account for 16.4% and 13.4% of all companies in the dataset respectively (see Figure 3.9). Despite these vague labels, it is evident the dataset skews towards high technology startups, as opposed to biomedical, agricultural, or other technologies (which do not make the Top 10).

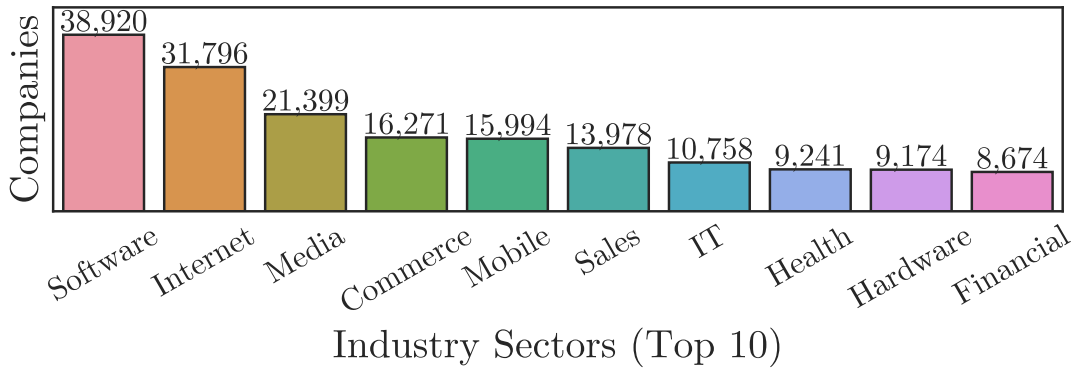


Figure 3.9: Companies grouped by industry sector. Industry labels are not mutually-exclusive. The 10 most common sectors are displayed.

3.3 Pipeline Creation

In the following section, we perform exploratory data analysis on our dataset and decide that a classification pipeline system could help us to address issues identified. The classification pipeline system is depicted in Figure 3.10.

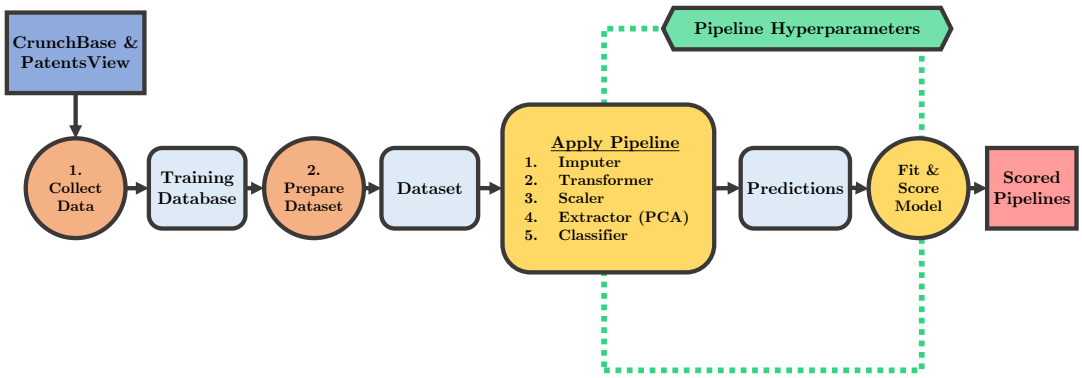


Figure 3.10: Pipeline creation overview. Grid search is performed across the pipeline hyperparameters to generate a variety of scored, candidate pipelines. Legend: dark blue square = input, orange circle = system component, yellow circle = process, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

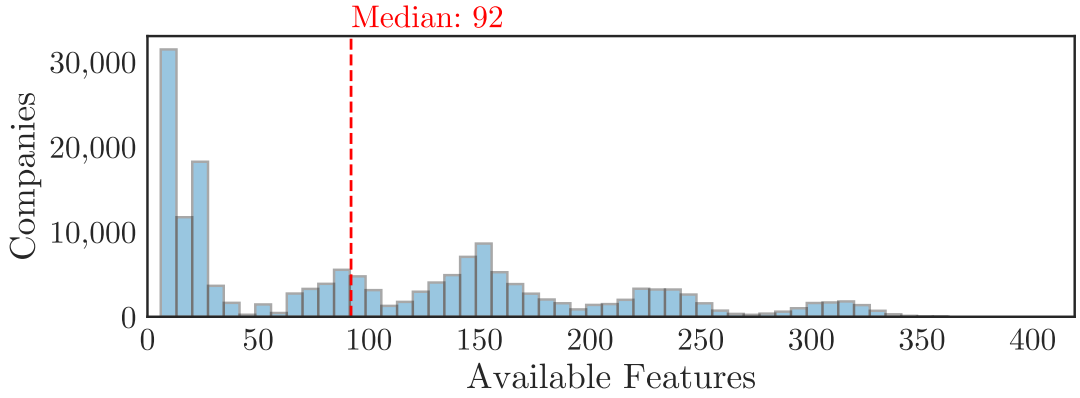
3.3.1 Exploratory Analysis

We performed exploratory data analysis on our dataset to assess what techniques would be suitable and the need for any further pre-processing. We identified dataset issues that include sparsity, long-tailed distributions, imbalanced feature ranges, and non-orthogonality.

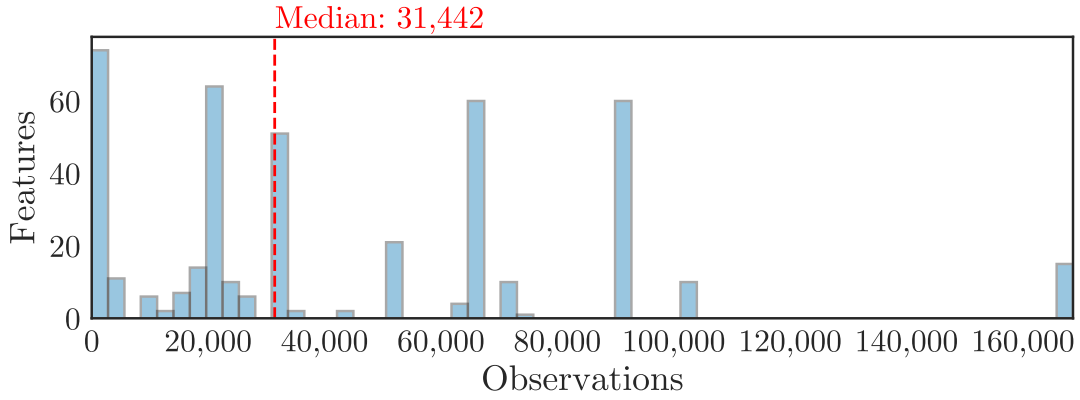
3.3.1.1 Sparsity

We explored the sparsity of the dataset. Sparsity is the distribution of null, zero or missing values. We expected the dataset to be highly sparse because CrunchBase is crowd-sourced. Figure 3.11 displays the distribution of features and observations in the dataset, with respect to each other. In Figure 3.11a we observe that many companies in the dataset have few available features (less than 50) and almost no companies have full feature sets. In Figure 3.11b we observe that few features have recorded observations for a large number of companies.

The multi-modal peaks of both figures suggest linkages between the availability of a subset of the features.



(a) Distribution of available features by company.



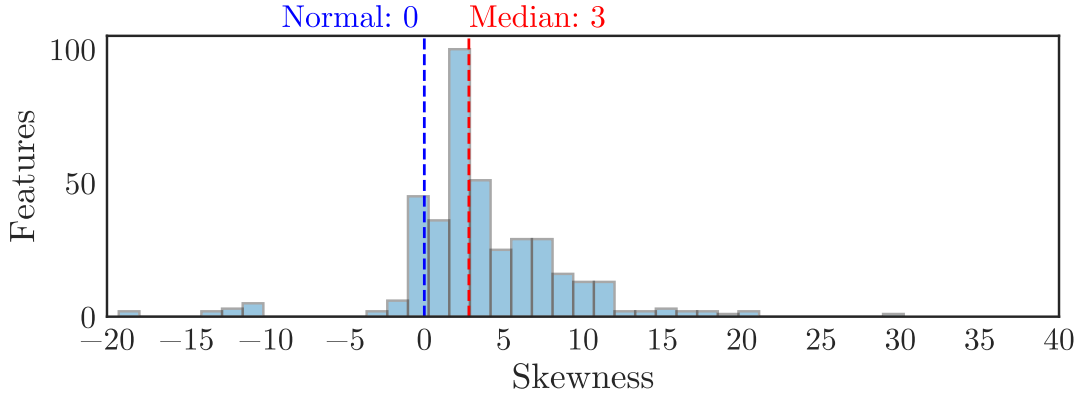
(b) Distribution of available observations by feature.

Figure 3.11: Density of features in our dataset.

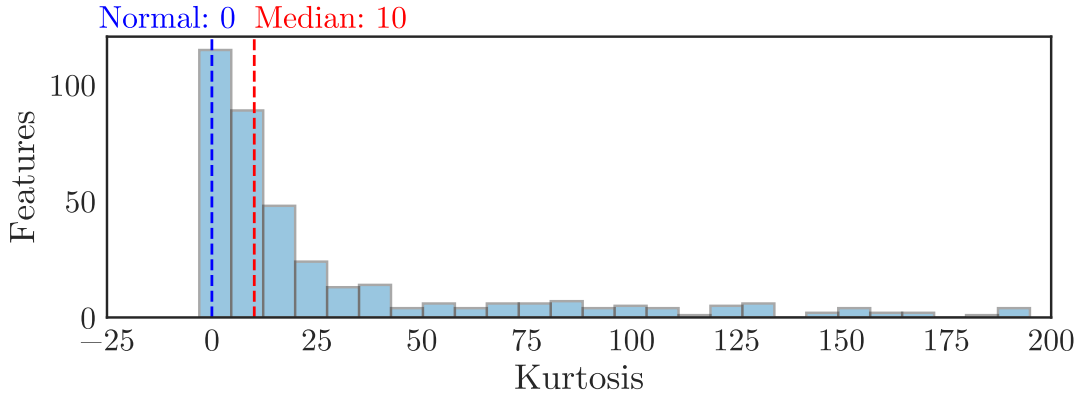
3.3.1.2 Normality

Next, we explored the normality of the dataset. Figure 3.12 shows the skewness and kurtosis of the features in our dataset. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. We consider a feature to be highly asymmetrical if its absolute skewness is above 1 and most of our features are more skewed than this cut-off. Kurtosis is a measure of the distribution of variance in a feature. If a feature has high kurtosis we call it 'long-tailed'. We use Fisher's measure of kurtosis, which has a normal value of 0. Our kurtosis distribution

suggests we have many extreme values (outliers) in our dataset. In combination, these results indicate that most features in our dataset are not normally distributed, but rather are positively-skewed, long-tailed distributions.



(a)



(b)

Figure 3.12: Normality of features in our dataset.

3.3.1.3 Scale

Next, we explored the scaling and range of each of our features. Figure 3.13 shows the Interquartile Range (IQR) of each feature. The distribution is extremely skewed in its original domain so we perform a log transformation to make the distribution easier to observe. Even the log-transformed distribution is highly skewed, which shows that our features have a wide range of magnitudes.

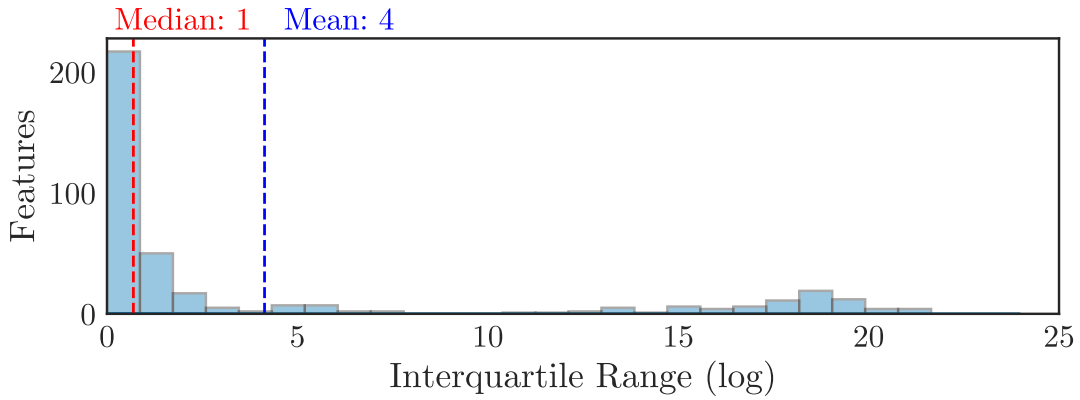


Figure 3.13: Distribution of interquartile ranges (log-transformed) in our dataset.

3.3.1.4 Orthogonality

Finally, we explored the orthogonality of our features: the extent to which the variance of our features is unrelated. We examined the distribution of pair-wise inter-correlations between our features, as depicted in Figure 3.14. We use two correlation metrics: Pearson and Spearman. Pearson is more commonly used, but Spearman is a ranked metric and may more accurately reflect our non-normal feature distributions [26]. Although most features have small inter-correlations ($\tilde{6}0\%$ below 0.2), there are still some that are highly correlated, so it might be efficient to remove these features using unsupervised feature extraction.

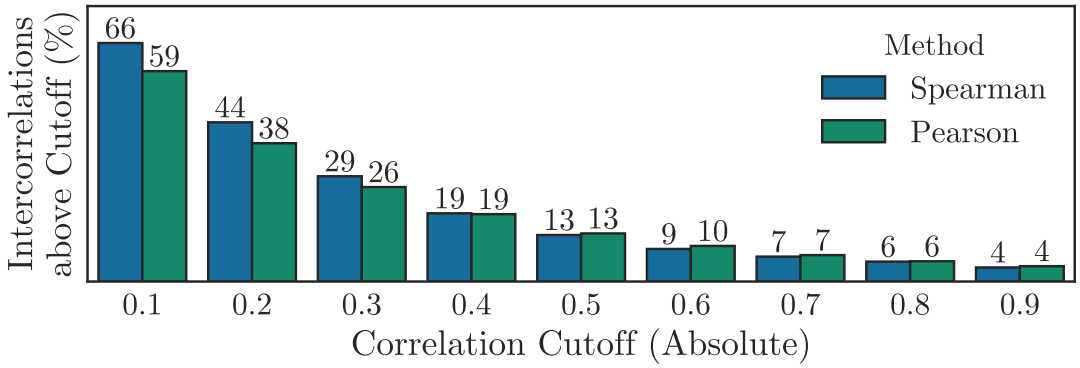


Figure 3.14: Distribution of inter-correlations in our dataset.

3.3.2 Hyperparameter Evaluation

To address the issues identified in the previous section, we developed a classification pipeline using the popular Python-based machine learning library Scikit-learn [27]. The classification pipeline construct allows us to search across hyperparameters at each step in the pipeline (e.g. imputation strategy, the number of components extracted in Principal Component Analysis (PCA), see Appendix D for the full hyperparameter list). The following section explores the evaluation of the pipeline hyperparameters against an indicative dataset generated from our training database. The indicative dataset is composed of a feature vector from April 2012 and an outcome vector from April 2014.

3.3.2.1 Imputation

After reviewing the distribution of missing data, we decided to investigate imputation methods further. Common imputation strategies include replacing missing values with the mean, median or mode of each feature. Figure 3.15 shows the distribution of mean, median and modes for each feature in the dataset. We apply a log-transformation to this figure to make the highly-skewed distribution easier to observe. There is minimal variance between the mean, median and mode of the features. For the majority of features, all three measures of central tendency are equal to zero. This finding resolves the issue of distinguishing missing data from negative observations because, following imputation, all of these data points will map to zero. Figure 3.16 shows the performance of the imputation strategies. All three imputation strategies produce similar results.

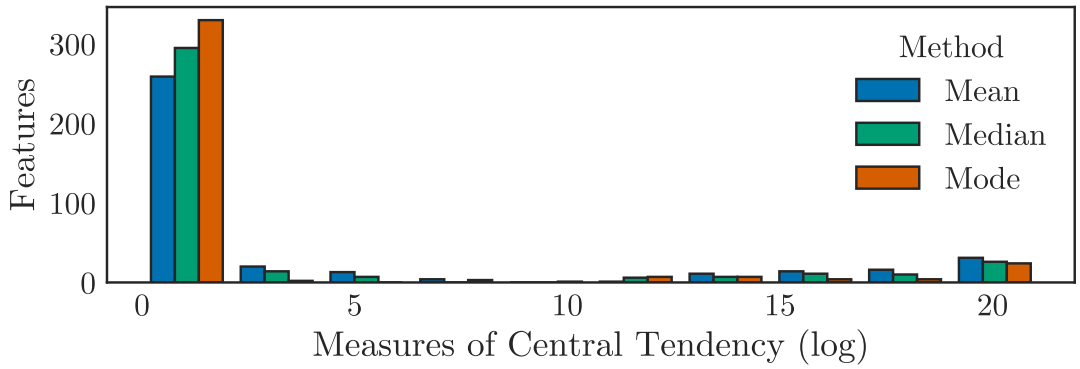


Figure 3.15: Distribution of measures of central tendency (mean, median and mode) in our dataset.

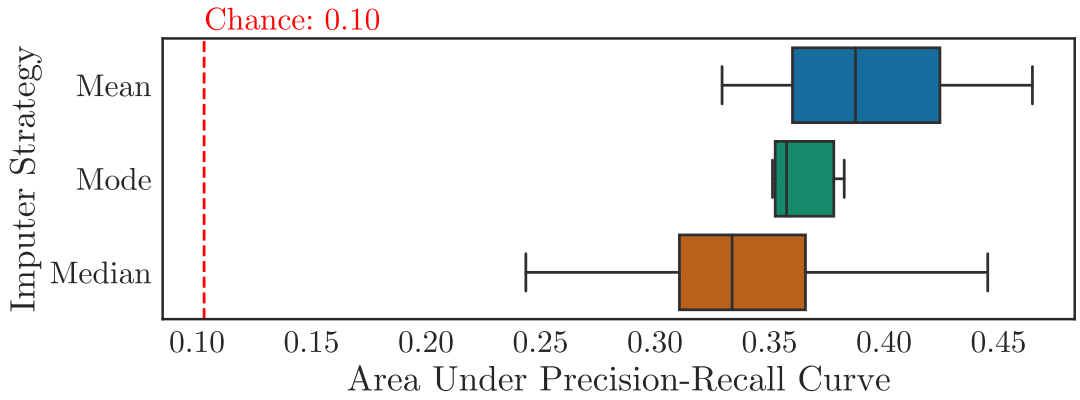


Figure 3.16: Area under Receiver Operating Characteristic (ROC) for different imputation strategies. Imputation strategies include replacing missing values with the most frequent (mode), median and mean value of each respective feature.

3.3.2.2 Transformation

While the classification algorithms we identified in the previous chapter are robust to violations of normality, it may be beneficial to transform the data if the feature distributions are extreme. Figure 3.17 shows one of the key features, Total Funding Raised, under different transformations. Like many features in our dataset, the distribution of Total Funding Raised is highly skewed. The log transformation reduces this skewness, and square root transformation also reduces this skewness (to a lesser extent). Figure 3.18 shows the performance of these transformation functions. Both functions provide a small improvement, with square root narrowly best.

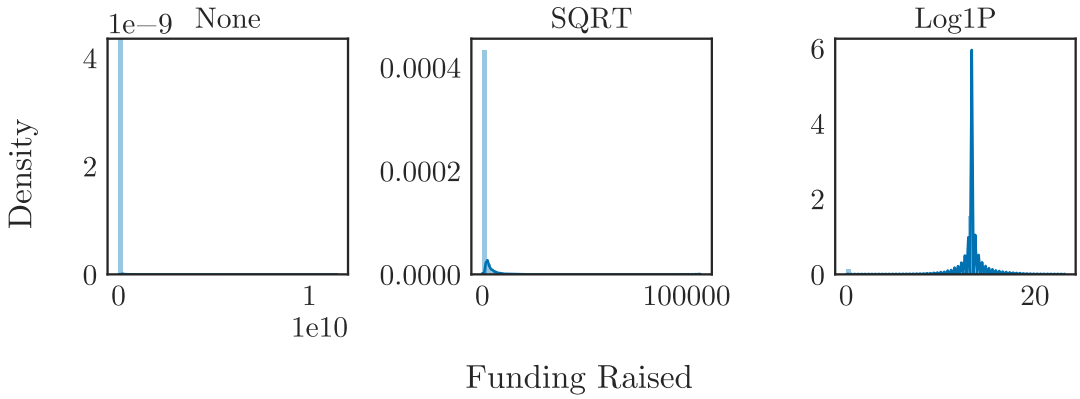


Figure 3.17: Funding raised transformed by functions.

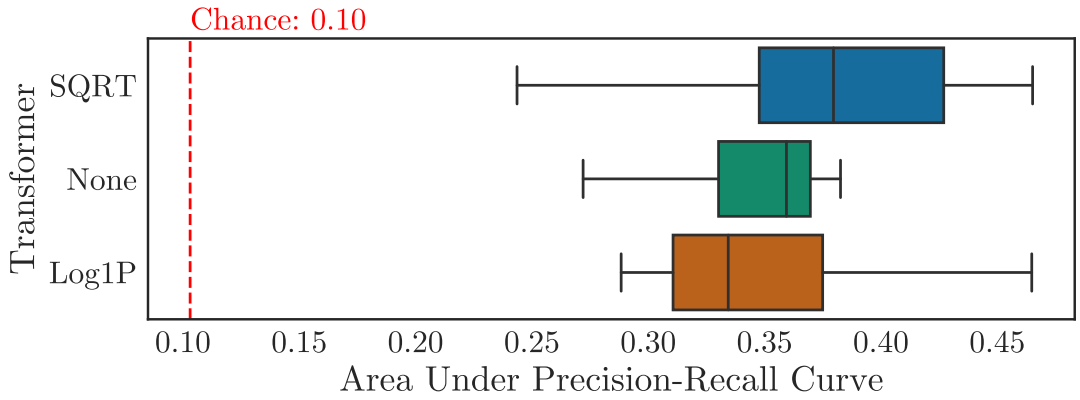


Figure 3.18: Area under ROC for different transformation functions. Transformations include: None (identity transformation), Log1p (natural logarithm of one plus the input array, element-wise), and SQRT (the square root of the input array, element-wise).

3.3.2.3 Scaling

Standardisation of datasets is a common requirement for many feature extraction methods and machine learning estimators. Scikit-learn provides three primary scaling functions: StandardScaler, RobustScaler and MinMaxScaler. RobustScaler is intended to alleviate the effect of outliers while MinMaxScaler is designed to preserve zero entries in sparse data - both of these are relevant properties for the dataset. Figure 3.19 shows the performance of these scaling functions. No scaling functions outperform the null condition which may be because the transformer already performs some scaling in an earlier step.

3.3.2.4 Extraction

Feature extraction reduces high-dimensional data into lower-dimensional data in such a way that maximises the variance of the data. The most common approach to dimensionality reduction is PCA. PCA is a technique which takes a set of vectors and finds an uncorrelated coordinate system in which to represent these vectors [28]. The magnitude of each eigenvector (eigenvalue) is displayed in Figure 3.20. The first ten components capture the majority of explained variance, and the eigenvalues drop below one by 100 components – this suggests that these are reasonable values for further hyperparameter search. Figure 3.21 shows the ROC for different numbers of extracted components. All curves produce similar classification results (within the margin of error) which implies that we should

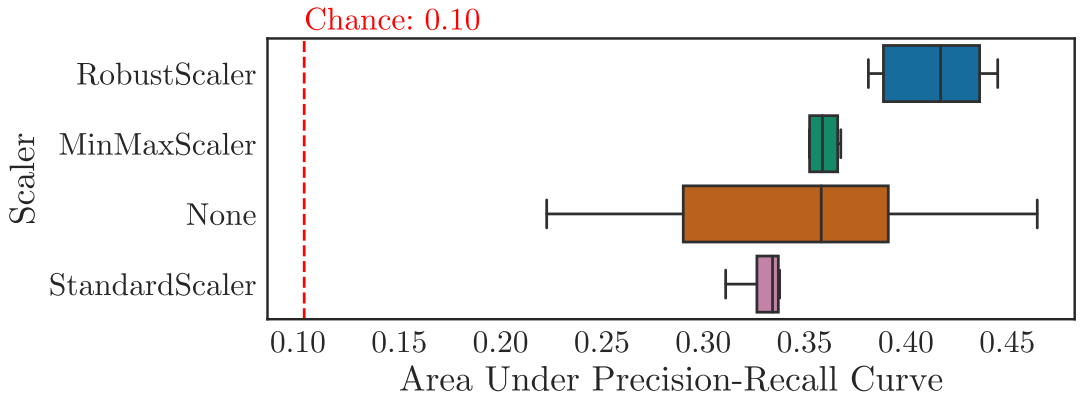


Figure 3.19: Area under ROC for different scaling functions. Scaling functions include: None, StandardScaler (mean: 0, variance: 1), RobustScaler (median: 0, IQR: 1) and MinMaxScaler (min: 0, max: 1).

extract between 1-20 components because it will provide us with more efficient computation.

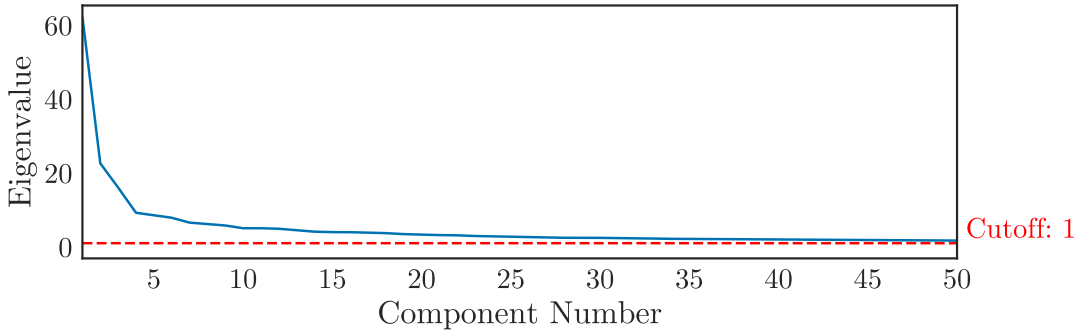


Figure 3.20: Eigenvalues extracted from PCA model. Horizontal line drawn at an Eigenvalue of 1 – this theoretically represents the 'contribution' of one original feature and is commonly used as an approximate threshold for included components.

While PCA is efficient at reducing features, the resultant components are not interpretable. However, analysis of 400+ individual features is also difficult to interpret. A compromise is to group features using our conceptual framework. We apply an approach that weights each feature to maximise the inter-correlations within each group. We use Spearman Correlation which is more robust than Pearson Correlation to extreme skewness [26]. Figure 3.22 displays the inter-correlations between each grouped factor. Overall, there is little correlation be-

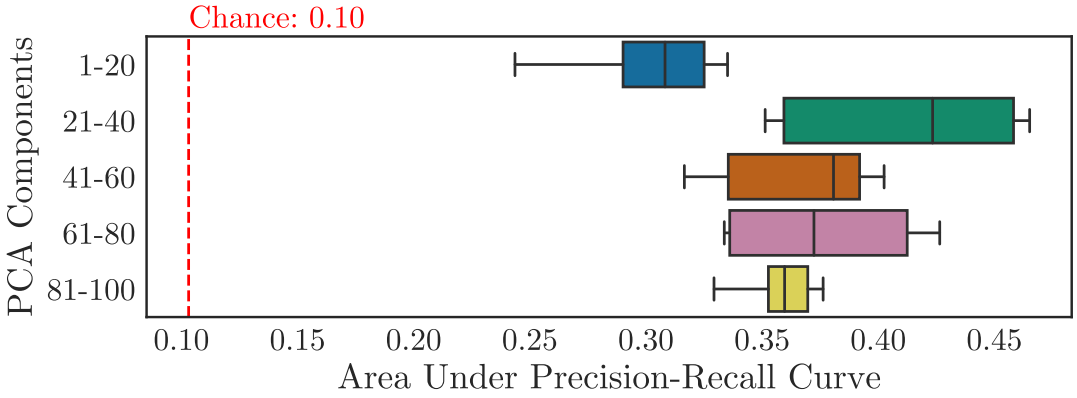


Figure 3.21: Area under ROC for different number of extracted components from PCA. Curves have been grouped by the quotient of the number of components divided by 20 to result in five ordered groups (e.g. Range [0, 19] becomes 0).

tween these grouped factors. This finding is promising because it implies that each group provides unique information. There are some minor correlations: Advisors are somewhat correlated with Founders and Investors, and Economic factors are somewhat correlated with Executives and Funding factors.

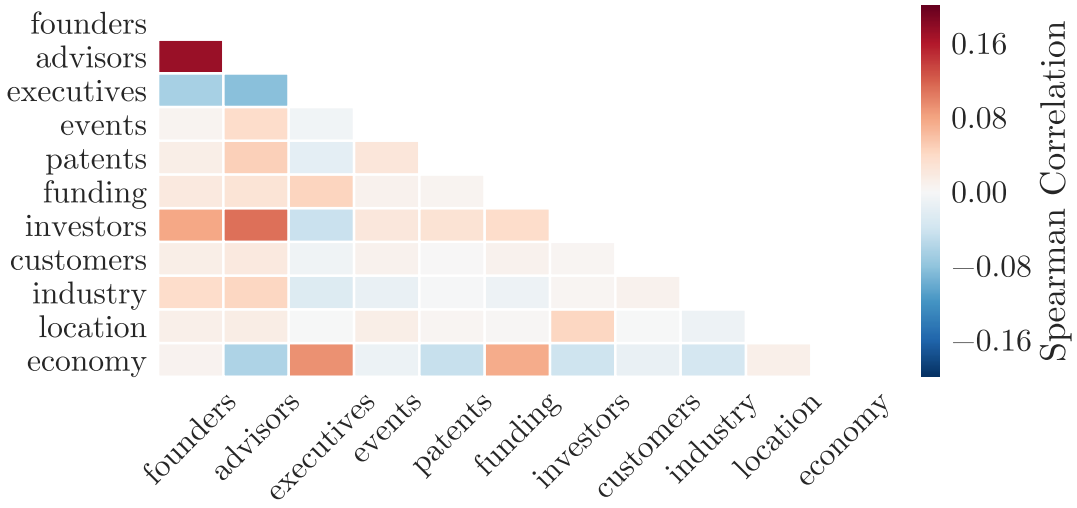


Figure 3.22: Inter-correlations of each factor from conceptual framework. Spearman ranking correlation is used. Individual features are grouped by applying weights that maximise the inter-correlations within each group from our conceptual framework (see Figure 3.2).

3.3.2.5 Classification Algorithms

The literature review we performed in the previous chapter identified common supervised classification algorithms potentially suitable for application to Venture Capital (VC) investment screening. Our analysis suggested that Random Forests were most likely to provide an optimal trade-off between predictive power, interpretability and time taken. We empirically tested each of these classifiers and compared their performance against a range of metrics, as displayed in Table 3.2. We report maximum and median scores so as not to penalise algorithms that have unfavourable hyperparameter search spaces.

	AUC PRC		AUC ROC		F1		MCC		Fit Time (s)	
Classifier	Median	Max	Median	Max	Median	Max	Median	Max	Median	75th
LR	0.417	0.465	0.675	0.710	0.339	0.358	0.255	0.288	7.3	412.7
RF	0.376	0.465	0.619	0.709	0.332	0.360	0.271	0.288	68.3	69.0
DT	0.388	0.429	0.651	0.659	0.305	0.314	0.212	0.224	15.3	16.8
NB	0.354	0.367	0.623	0.638	0.303	0.321	0.212	0.239	8.6	26.8
KNN	0.335	0.353	0.532	0.565	0.131	0.226	0.137	0.210	8.5	20.8
ANN	0.320	0.335	0.517	0.523	0.072	0.096	0.111	0.140	9.1	21.0
SVM	0.233	0.244	0.503	0.504	0.014	0.017	0.038	0.045	29.0	29.0
Total	0.357	0.465	0.623	0.710	0.300	0.360	0.209	0.288	15.3	29.0

Table 3.2: Overview of classification algorithm performance. Algorithms are: NB = Naive Bayes, LR = Logistic Regression, KNN = K-Nearest Neighbours, DT = Decision Trees, RF = Random Forests, SVM = Support Vector Machines, ANN = Artificial Neural Networks.

We take a closer look at the Precision-Recall (PR) curves for each classifier in Figure 3.23. While all classifiers perform better than chance, Logistic Regressions and Random Forests come out ahead, and Support Vector Machines and Artificial Neural Networks appear to underperform. Examining the cross-validated learning curves for each classifier (Figure 3.24), we see that Naive Bayes, Logistic Regression, Artificial Neural Networks and Support Vector Machines quickly converge, whereas Decision Trees, Random Forests and K-Nearest Neighbours require more observations to converge. This suggests that Random Forests may do better in final testing (which will not be cross-validated) and in the future as the dataset grows.

3.4 Pipeline Selection

In this step, we evaluate the best pipelines from the previous step over different dataset slices. This process, depicted in Figure 3.25, ensures our final pipeline

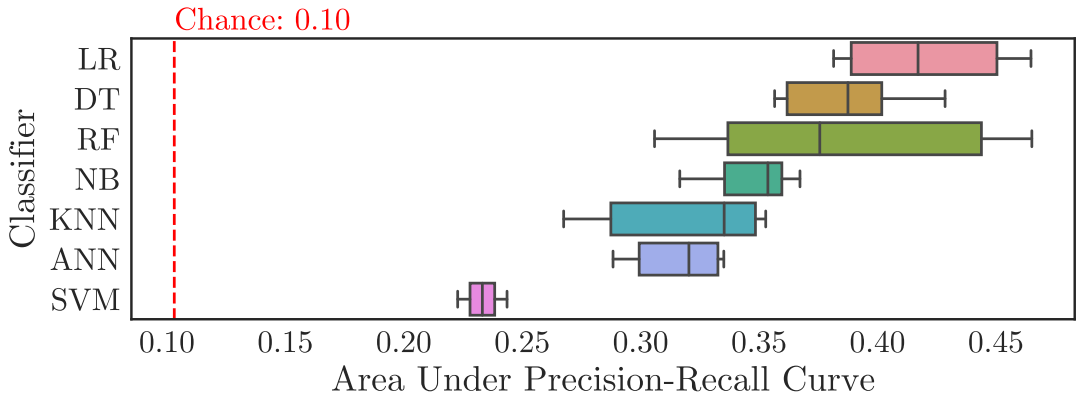


Figure 3.23: Area under ROC for different classification algorithms. All algorithms are implementations from the Sci-kit learn library. Algorithms are: NB = Naive Bayes, LR = Logistic Regression, KNN = K-Nearest Neighbours, DT = Decision Trees, RF = Random Forests, SVM = Support Vector Machines, ANN = Artificial Neural Networks.

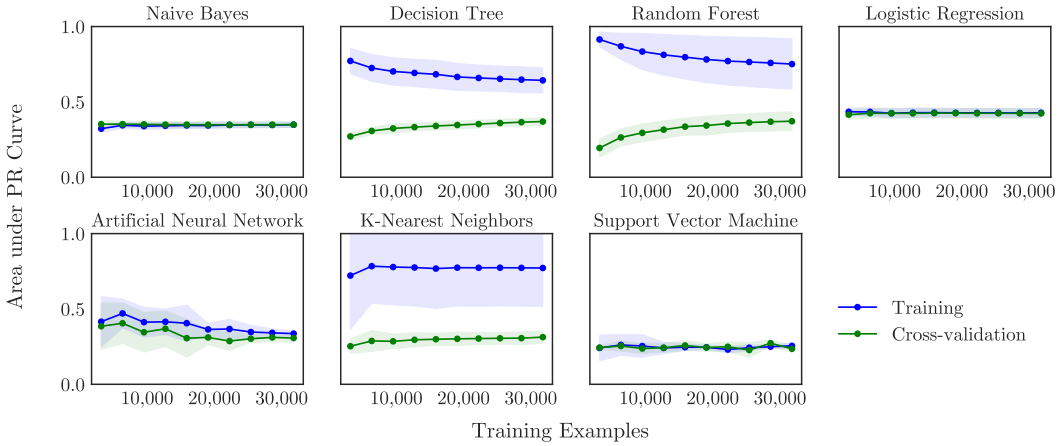


Figure 3.24: Learning curves by classification algorithms.

is robust in its performance over time. We aggregate the results for each finalist pipeline across these dataset slices and rank the finalist pipelines on their overall performance. Finally, we select the best pipeline.

3.4.1 Evaluation Metrics

Next, we decided how to select finalist pipelines that we can evaluate further. There are a variety of metrics used to assess binary classification algorithms.

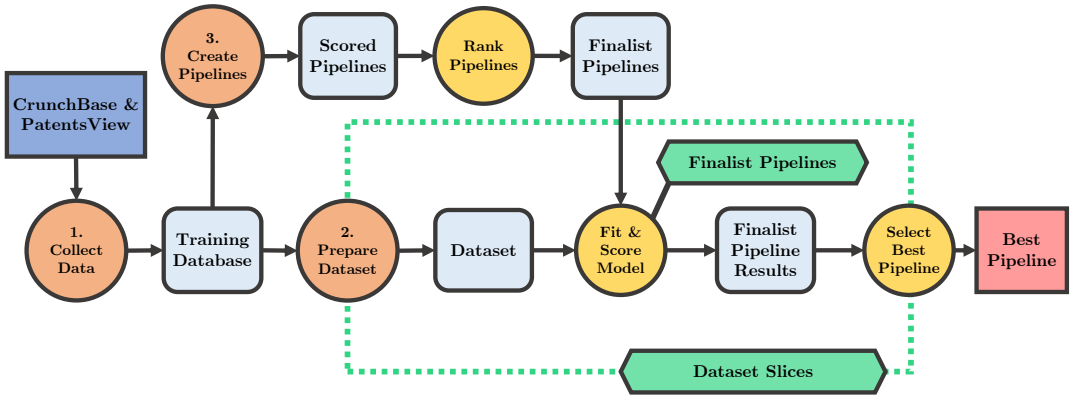


Figure 3.25: Pipeline selection overview. Legend: dark blue square = input, orange circle = system component, yellow circle = process, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

Accuracy is rarely used in practice because it gives misleading results in the case of imbalanced classes. Receiver Operating Characteristic (ROC) curves are commonly used, which show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. The area under these curves gives a standardised result across a spectrum of decision thresholds. Precision-Recall (PR) curves are similar to ROC curves but instead map the trade-offs between precision and recall. They are less commonly used than ROC curves but have been shown to produce more accurate results for imbalanced classes than ROC curves [29]. We decided to proceed with PR curves because our dataset is highly imbalanced. We will also use this metric to rank our finalist pipelines.

3.4.2 Finalist Pipeline Evaluation

Our hypothesis is that the performance of our pipelines may vary with respect to the dates of our datasets. To evaluate this hypothesis, first, we explored variance between the pipelines on aggregate against the slice dates, presented in Figure 3.26. There is no relationship observed between slice date and score.

Next, we study the variance within the individual pipelines, presented in Figure 3.27. Although there is a positive correlation between the pipelines initial ranking and their scores, there are deviations. For example, the top-ranked pipeline from the first stage has a lower median score than the second-ranked pipeline. These results suggest that the top 3-5 pipelines should be evaluated in

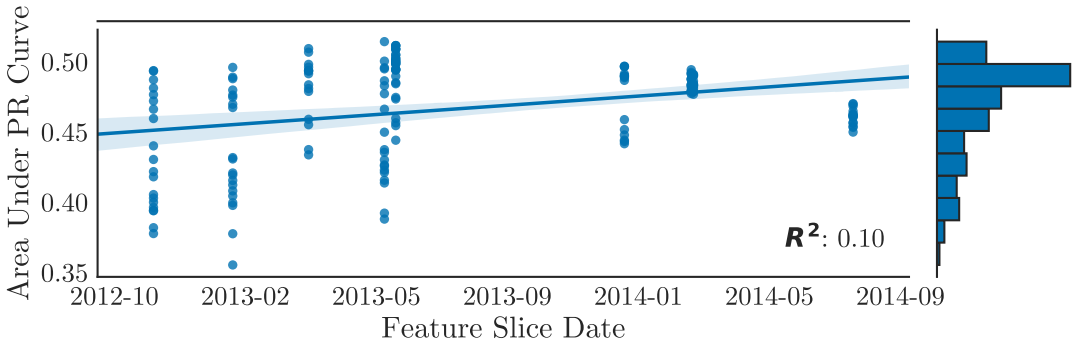


Figure 3.26: Pipeline performance by slice date.

this manner to ensure that the best pipeline is selected. We describe the chosen pipeline in Table E. We adopted this pipeline for our following experiments.

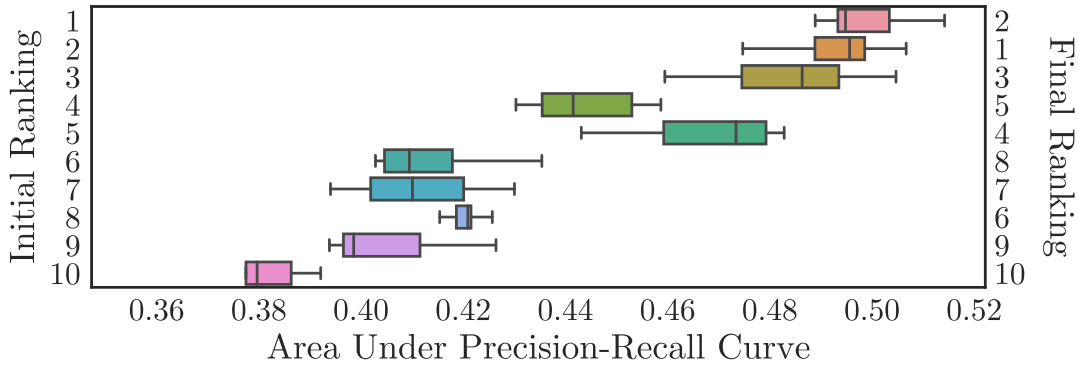


Figure 3.27: Overview of finalist pipeline performance.

3.5 Model Fit and Prediction

Finally, our system applies the best classification pipeline to a feature vector from a held-out test database to generate a model and a set of predictions, as shown in Figure 3.28. We evaluate the accuracy of the models produced by our system in the next chapter.

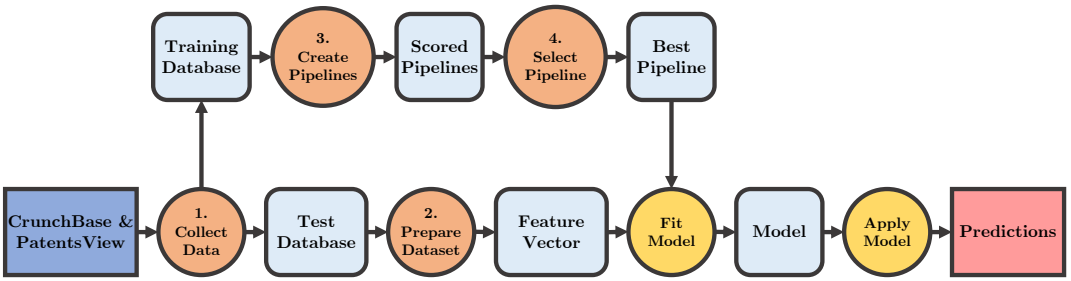


Figure 3.28: Model fit and prediction overview. Legend: dark blue square = input, orange circle = system component, yellow circle = process, light blue rounded square = intermediate, red square = output.

CHAPTER 4

Evaluation

We believe it is possible to produce an improved Venture Capital (VC) investment screening system. In Chapter 3, we described the design and development of such a system. Our system identifies startup companies likely to receive additional funding or exit in a given forecast window. In this chapter, we perform a series of experiments to evaluate our system against criteria of practicality, robustness and versatility.

1. Practicality. We designed our system to require minimal user input. Therefore, by virtue of its design, the proposed system is more efficient than current systems. We also explored the time profile of our system. An indicative implementation of our system takes 46 hours to run.

2. Robustness. We observed minimal variance in the performance and types of models generated by our system across training sets of various dates. We also evaluated the learning curves of our system and identified that our system is likely to adapt and perform better as the data sources grow over time.

3. Versatility. We assessed our system's ability to perform a variety of investment prediction tasks. Tasks included predicting companies at different developmental stages, for different target outcomes, and over different forecast windows. Our system performs best for longer forecast windows (up to 4 years) and companies at later developmental stages (e.g. Series B, C).

4.1 Experimental Design

In this chapter, we evaluate models generated by our system while varying a number of other factors. This evaluation process is depicted in Figure 4.1. The optimised pipeline is fit to a training dataset to generate a model. The model is

applied to a test feature vector to produce predictions. We scored these predictions against truth values derived from a held-out test database collected in April 2017. This process is performed multiple times to evaluate the system against the three criteria derived from our literature review: practicality, robustness and versatility. The configuration of the system during our experiments is detailed in Appendix E.

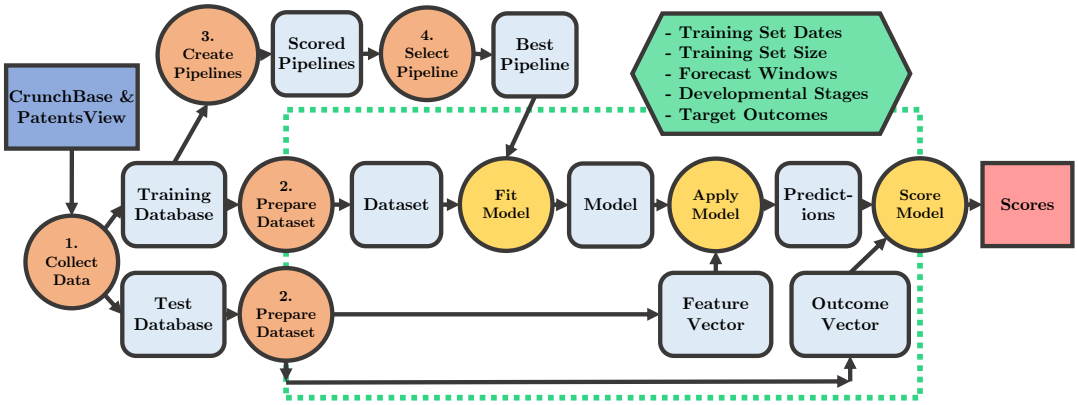


Figure 4.1: Pipeline evaluation overview. Training and test datasets are generated according to the experimental configuration: varying with respect to training set dates, training set size, forecast window, developmental stage, and target outcomes. Legend: dark blue square = input, orange circle = system component, yellow circle = process, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

4.1.1 Baseline Analysis

Before we evaluated our system, we performed preliminary analyses to determine the baseline trends and distributions of company outcomes in our database.

First, we explored company outcomes by forecast window. We applied the same system of reverse-engineering time slices that we used in previous experiments on robustness, but this time we varied the time passed between our feature vector and outcome vector (i.e. the forecast window). We created datasets from feature and outcome vectors of each year from 2012-2016 and explored the proportion of companies that raised additional funding, were acquired or had an Initial Public Offering (IPO).

Figure 4.2 shows how company outcome varies with respect to the forecast window (time between the observed features and the measured outcome). We observe a positive relationship between length of forecast window and company

outcome. For additional funding rounds, this relationship disappears after three years. This finding implies that companies that do not raise additional funding rounds over a three year period are unlikely to raise further funds at all. We do not observe this effect for acquisitions or IPOs which implies there is either greater variability in the time it takes companies to exit or the lead-time to exit for most companies in our dataset is longer than four years. Few companies exited (1.2%) or raised funds (5%)over a period of fewer than two years which suggests we should focus our experimentation on forecast windows of 2–4 years.

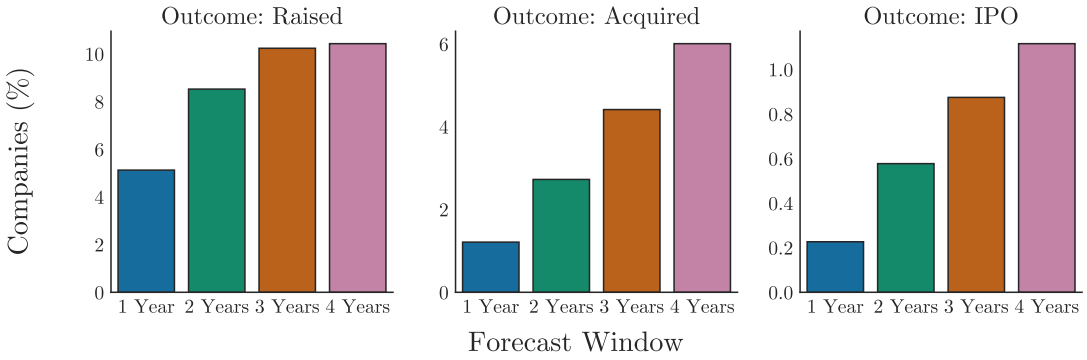


Figure 4.2: Outcomes by forecast window.

We also explored how company outcomes vary with respect to development stage, shown in Figure 4.3. We see a broad positive relationship between developmental stage and likelihood of further funding rounds and exits. There is a high proportion of later stage companies in our dataset that do not receive additional funding or exit over four years. Only a small proportion of companies in our dataset are reported to be Closed. We can guess that many companies that do not raise or exit may have closed or down-sized. We also observe considerable variation in the outcomes of companies of different developmental stages. For example, only a small proportion of companies perform an IPO before Series C stage. We decided to investigate how our system predicts outcomes for each developmental stage independently, as well as in aggregate.

4.1.2 Evaluation Metrics

While Area under the Precision-Recall (PR) Curve was used to guide the development of our system, in evaluation of our system's performance we primarily use F1 Scores. An F1 score is the harmonic mean of recall and precision at points on the PR curve. The harmonic mean is an alternative to the arithmetic mean that tends strongly towards the smallest element in a set. Accordingly, F1 Scores

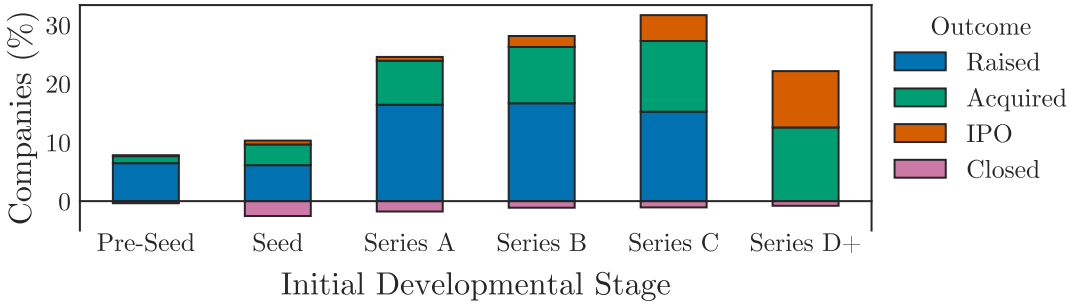


Figure 4.3: Outcomes by developmental stage.

are highly sensitive to trade-offs between recall and precision. The Area Under Curve (AUC) measure provides an overall evaluation of a classification system, whereas the F1 Score evaluates a set of predictions. For investment screening, we are more sensitive to classification performance for the positive class (companies that have been successful in raising further funding or achieving an exit), so hereafter, when we refer to F1 Score, we refer to the F1 Score for this class alone. We also present Matthews Correlation Coefficient (MCC) in some of our analyses. MCC is a measure of the correlation between the observed and predicted binary classifications. The MCC should produce similar results to an F1 Score that incorporates performance across both positive and negative classes.

4.2 Practicality

The Venture Capital (VC) industry requires more efficient forms of investment analysis, particularly in surfacing and screening. VC firms currently perform these processes through referral, Google search, industry papers and manual search of startup databases. Our automated system is more efficient than these methods because it is designed to involve minimal user input. However, it is also important that our system is relatively time-efficient so that it can run frequently enough so that its predictions are always up-to-date and relevant. We assess the time profile of our system to determine whether it is practical for use in the VC industry.

An indicative time profile of the system is shown in Table 4.1. At the highest-level, this configuration of the program takes 46 hours to complete on a modern desktop PC. When we further break this time down by system component, the vast majority of time (84.8%) is taken up by the initial pipeline creation component. This time is due to the pipeline optimisation process – the model is fit and scored over 500 times on different classification algorithms and parameters.

Scoring takes a long time because, in this case, it also involves generating learning curves for reporting, which is another cross-validated process. In practice, we could reduce the running time of our system by removing logging and reporting processes and by scheduling components of the system to run independently. For example, the pipeline creation process takes a long time but is relatively robust to dataset changes in the short-term, so the system could potentially run this component on a more infrequent basis.

Function	Cycle (s)	Cycles (N)	Time (s)	Time (m)	Time (h)
Generate Dataset (CV)	1,800	1	1,800	30	0.5
Prepare Feature Dataset	1,200	1	1,200	20	0.3
Prepare Outcome Dataset	180	1	180	3	0.1
Merge Datasets	360	1	360	6	0.1
Finalise Dataset	60	1	60	1	0.0
Fit and Score Model[1]	265	525	139,125	2,319	38.6
Fit Model	15	525	7,875	131	2.2
Score Model	250	525	131,250	2,188	36.5
Subtotal: Create Pipelines			140,925	2,349	39.1
Get Finalist Pipelines	5	1	5	0	0.0
Generate Dataset (CV)	1,800	5	1,800	30	0.5
Fit and Score Model[2]	265	75	19,875	331	5.5
Select Best Pipeline	5	1	5	0	0.0
Subtotal: Select Best Pipeline			21,685	361	6.0
Generate Dataset (Training)	1,800	1	1,800	30	0.5
Generate Dataset (Test)	1,800	1	1,800	30	0.5
Fit Model	30	1	30	1	0.0
Make Predictions	5	1	5	0	0.0
Subtotal: Fit and Make Predictions			3,635	61	1.0
Total			166,245	2,771	46.2

Table 4.1: System time profile. All times are indicative based on averages from system logs. Notes: [1] Cycles involve 25 search iterations, 3 cross-validated folds, and 7 classification algorithms. [2] Cycles involve 5 finalist pipelines, 3 database slices, 3 cross-validated folds.

4.3 Robustness

An improved Venture Capital (VC) investment screening system must be robust to changes over time. VC firms have concerns that models trained on historical

data will not predict future trends and activity. These concerns are a key barrier to the adoption of automated systems by the VC industry [14]. Therefore, it is critical that our system is shown to be robust in its performance with respect to time so investors can rely on its predictions. Similarly, VC firms seek systems that are themselves robust to time and can adapt to new data sources and feature sets as they become available. In the following section, we evaluate our system based on its robustness to training set dates and training set sizes.

4.3.1 Training Set Date

A VC investment screening system should have minimal variance in its performance when training on datasets from different dates so investors can trust its ability to make future-looking predictions.

Figure 4.4 shows models trained on training sets sliced from various years and scored against key evaluation metrics. We held the forecast window constant at two years because we cannot test training sets from later years using long forecast windows, and this would skew our results. Variance across all evaluation metrics is low, with a slight improvement in performance for newer dataset slices which probably reflects that newer datasets provide more training data.

We explored the feature weights for each model in Figure 4.5. While there are some slight differences with respect to Executives, Investors, and Economic factors, the baseline distribution trend is similar across all training set dates. We discuss the distribution of these feature weights in more detail in a following section.

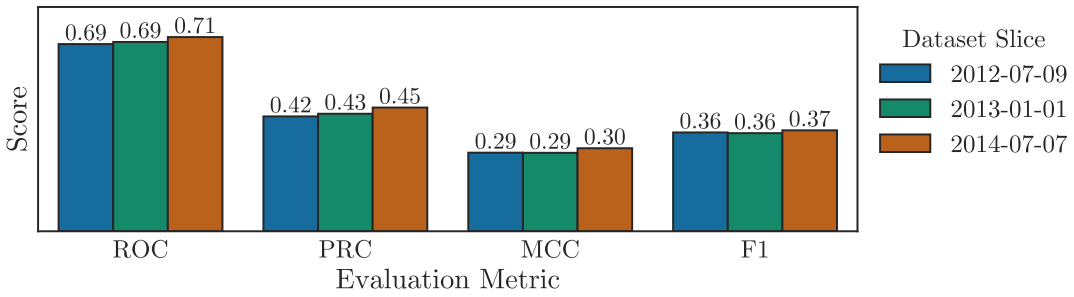


Figure 4.4: Performance by training set date.

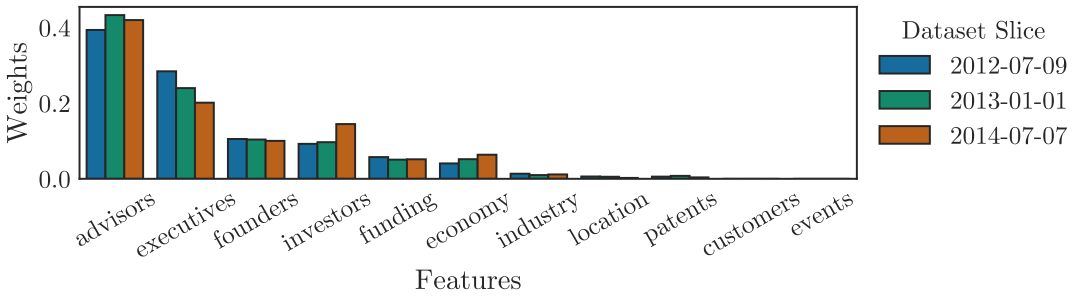


Figure 4.5: Feature weights by training set date.

4.3.2 Training Set Size

Learning curves allow us to evaluate how the bias and variance of a classification technique vary with respect to the amount of training data available. We investigated learning curves for our system to determine whether our system's performance has potential to improve as its data sources grow. We sampled our training sets five times at different fractional rates. The rate of convergence (or divergence) of our training and cross-validation curves can indicate whether our classification pipeline is over- or under-fitting our data for various sizes. We investigated these learning curves with respect to forecast window, target outcome and developmental stage.

Figure 4.6 shows the learning curves for forecast windows of 2–4 years against a combined target outcome and companies from all developmental stages. The maximum number of training examples is negatively related to the length of the forecast window because newer datasets have more examples. For a forecast window of four years, the curves have converged, whereas for shorter forecast windows there still seems to be some benefit to additional training examples. Much of the testing score improvement comes in the first 20,000 training examples, which suggests this pipeline is approaching optimal performance.

We can measure investment success by various target outcomes. We have evaluated previous learning curve plots against a combined outcome of either raising funds, being acquired or having an Initial Public Offering (IPO). We termed this "Extra Stage". We see that the efficiency of our system varies with respect to these outcomes in Figure 4.7. Predicting whether a company raises an extra round is the least data-intensive outcome, as it converges quickly. In comparison, the plot for predicting company exits does not converge. Our model has most difficulty predicting IPOs, probably because these are rare events in our dataset.

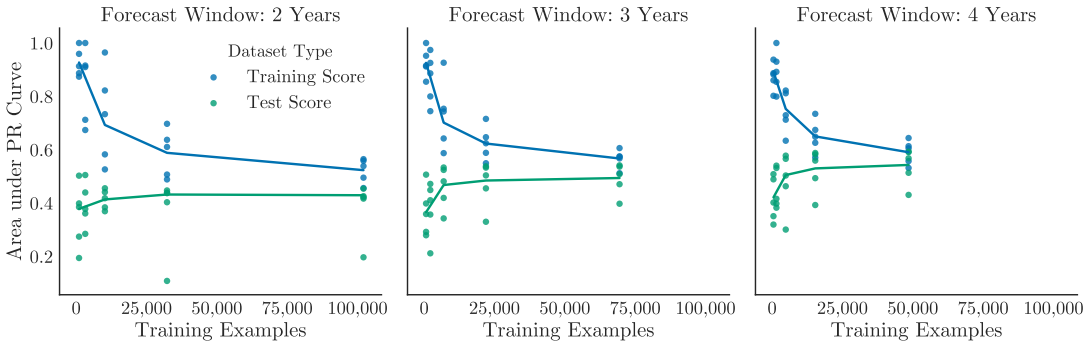


Figure 4.6: Learning curves by forecast window.

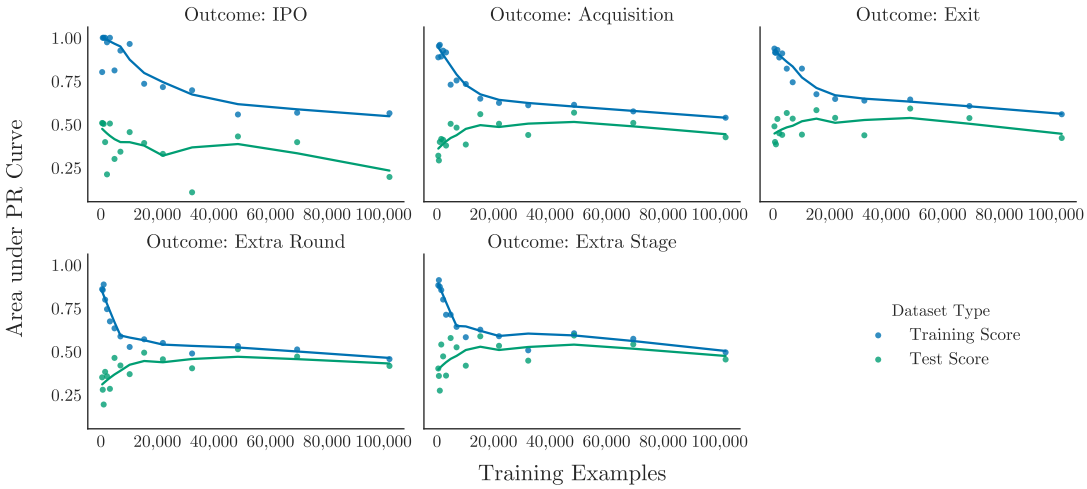


Figure 4.7: Learning curves by target outcome.

Finally, Figure 4.8 shows the learning curves of models fitted independently to companies of different developmental stages, for a forecast window of four years and a combined target outcome. We observe significant variance in the learning curves for various developmental stages. In fact, it appears that developmental stage is the dominating factor across all of the learning curve plots. Pre-Seed, Seed and Series A, which make up the majority of the dataset, have converged or are at near-convergence at relatively low scores. Companies at these stages probably require more features or more complicated classification algorithms (e.g. Artificial Neural Networks) to improve their performance further. This finding may be related to the observation that companies in earlier developmental stages have the most missing features in our dataset. Unlike companies at earlier stages, the learning curves of Series B, C and D+ imply that more training examples will improve the performance of these models.

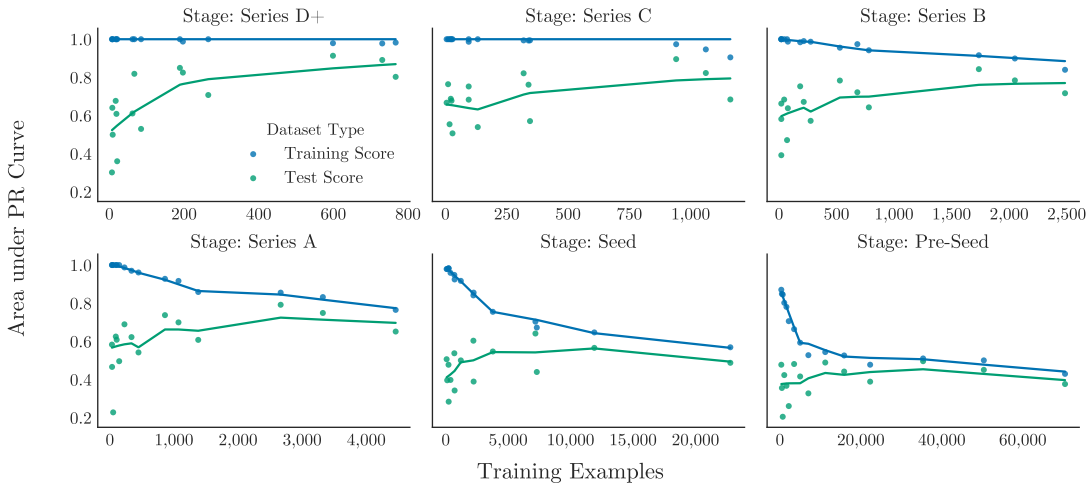


Figure 4.8: Learning curves by developmental stage.

4.4 Versatility

Our system must be consistently accurate at identifying a variety of high-potential investment candidates. We evaluated the systems' versatility based on its ability to predict over different forecast windows (e.g. 2–4 years), for target companies at different developmental stages (e.g. Seed, Series A, etc.), and for different target outcomes (e.g. predicting additional funding rounds, being acquired, having an Initial Public Offering (IPO), or some combination thereof).

4.4.1 Forecast Window

A forecast window is the period between when a prediction is made and when that prediction is evaluated (i.e. a prediction made in 2014 on whether a company would exit by 2017 is a forecast window of three years.) The Venture Capital (VC) industry raises funds with fixed investment horizons (3–8 years) [3], so time to payback is a key component of VC investment decision-making and portfolio management. It is important we understand how the models and predictions produced by a VC investment screening system vary with respect to the length of these forecast windows.

Figure 4.9 shows model performance across a range of metrics, grouped by forecast window. We observe little difference in Area under the Receiver Operating Characteristic (ROC) curve across the forecast windows. However, across all three other metrics, there is a positive relationship between length of forecast window and model performance. The F1 Score shows the greatest improvement

in performance over time (52.7%), compared to Area under the Precision-Recall (PR) curve (34.1%) and Matthews Correlation Coefficient (MCC) (11.6%).

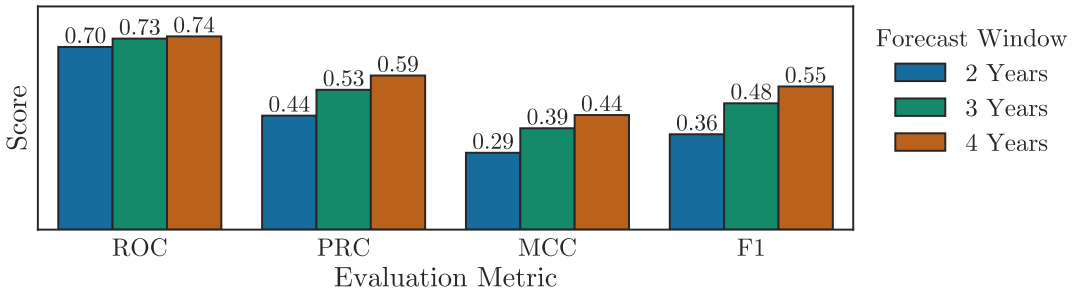


Figure 4.9: Performance by forecast window.

Figure 4.10 shows the standardised weights of features grouped using the conceptual framework proposed earlier in this paper, grouped by forecast window. First, we discuss the baseline distribution and then examine the variation in weightings with respect to forecast window. Factors related to advisors are the best predictor of startup investment success. Executives and founders features are also important factors and round out measures of human capital. The quality of investors that invest in a startup (assessed by their prior investments) is found to be more important than the quantum of investment raised by a startup. Local economy and industry features are weak predictors, as are customers and social influence (in this case, measured through participation at events). There is little difference between the weightings of each feature group with respect to forecast window. However, there are a few trends to point out: the importance of advisor factors increases over time, and the importance of executives and the broader economic factors decreases over time.

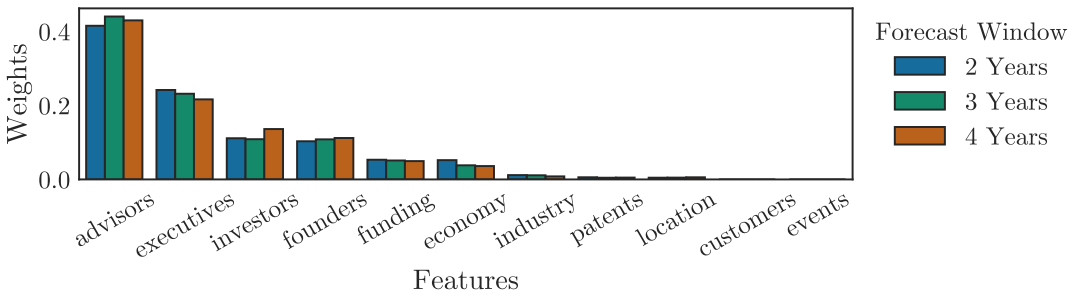


Figure 4.10: Feature weights by forecast window.

4.4.2 Development Stage

We can classify startups into developmental stages based on their external funding milestones. These milestones signal not only a change in the resources available to a startup, but also their functions and objectives, and in turn the type of investors that are interested in them as investment opportunities. In Chapter 3 we mapped the companies in our dataset to their developmental stages. In the following section, we evaluated how the system models and predicts the outcomes of companies at different developmental stages.

Figure 4.11 shows F1 Scores grouped by developmental stage and fit method. First, we examined the baseline distribution and then the variation in performance by fit method. Model performance has a positive relationship with developmental stage. The only deviation from this relationship is for Series D+. To understand this discrepancy better, we split our datasets into their developmental stages and fit the model onto each of these sub-datasets individually (i.e. each dataset contains a single developmental stage). This method results in a broad performance improvement which has the least impact on Pre-Seed and the greatest impact on Series D+ companies.

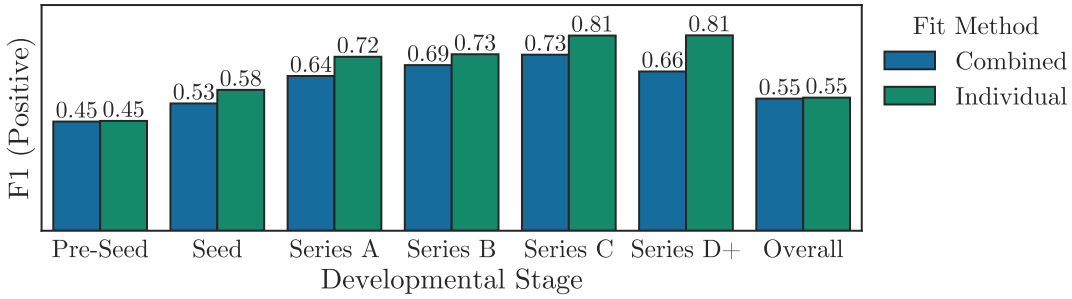


Figure 4.11: Performance by developmental stage.

Figure 4.12 shows the standardised weights of features, grouped by developmental stage. While we observe a similar trend to Figure 4.10, there is more variation in weights than was seen when grouped by forecast window. Advisors are more important to earlier stage companies than late stage companies, investor track record and reputation becomes important as companies approach an exit (Series D+), executive and founder experience are important in pre-seed companies, as is broader economic outlook.

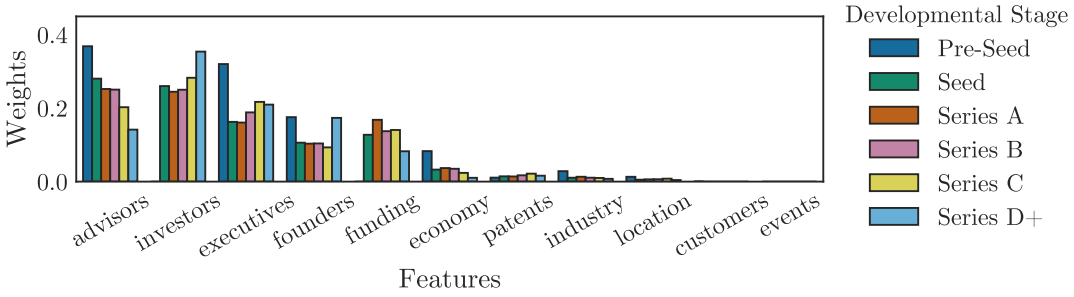


Figure 4.12: Feature weights by developmental stage.

4.4.3 Target Outcome

Ultimately, VC firms seek rare investments that will return their invested funds many times over within an investment horizon of their fund (3–8 years) [3]. Funds are only returned to VC investors when startups have liquidity events (IPO, Acquisition). However, recently, many companies that are considered successful are delaying their liquidity events and seeking later-stage private funding (e.g. Uber). In this case, whether a company has raised additional funding rounds may be used as a proxy for investment success. Unless otherwise specified, we performed our previous analyses against our base target outcome, Extra Stage (i.e. whether a company raises an additional funding round, is acquired or has an IPO). In the following section, we explore whether other target outcomes have an effect on our system's predictive power.

Figure 4.11 shows F1 Scores grouped by target outcome and forecast window. First, we examine the baseline distribution and then the variation in performance by forecast window. Our model is most accurate at predicting extra funding rounds and worst at predicting IPOs. As we observed in Figure 4.9, there is a positive relationship between length of forecast window and model performance. This relationship has a similar magnitude across all target outcomes except for IPOs which improve more when we increase the forecast window from two to three years.

Figure 4.14 shows the standardised feature weight distribution, grouped by target outcome. Models of target outcomes produce considerable variance in feature weights. Exit and Acquisition have similar feature weights. Investors, Executives and Founders are key features for Exits and Acquisitions. In comparison, IPOs have greater weighting towards Funding, Advisors and the Broader Economy. Extra Round and Extra Stage have similar feature weights and are most strongly related to Advisors and Executives.

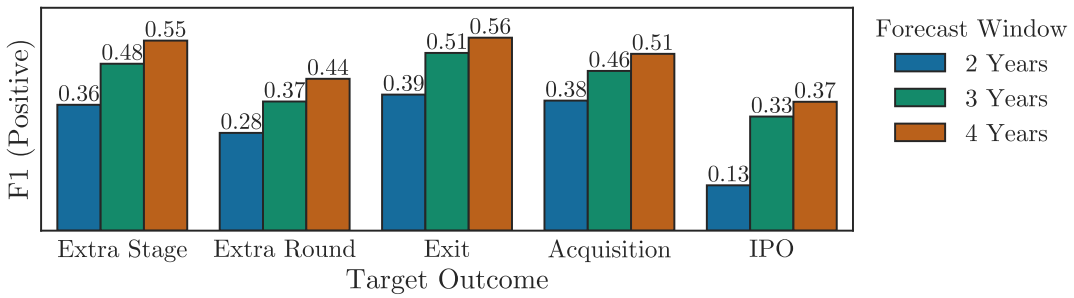


Figure 4.13: Performance by target outcome.

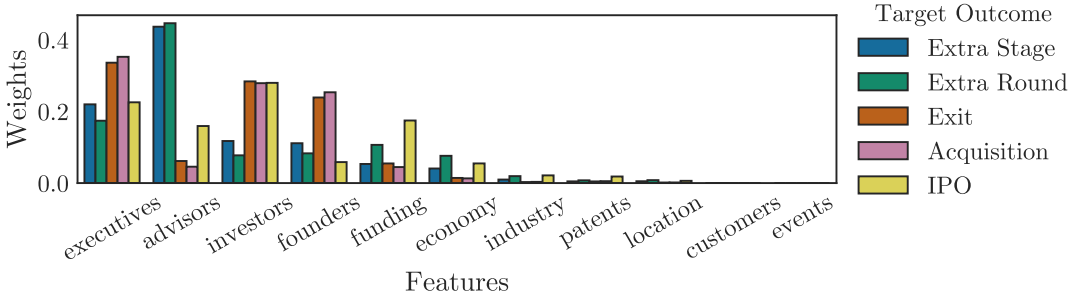


Figure 4.14: Feature weights by target outcome.

CHAPTER 5

Discussion

In previous chapters, we described how we designed and evaluated a novel data mining system for automated Venture Capital (VC) investment screening. In this chapter, we discuss the merits of this project with respect to our system's design, our system's performance, and its contributions to theory on VC investment.

1. System Design. We developed a data-mining system that provides automated VC investment screening. Our system leverages a comprehensive feature set from CrunchBase and PatentsView. We developed a pipeline optimisation system to address limitations in these data sources and adapt to their changes over time. Our system is semi-autonomous and designed so that the VC industry can adopt it with minimal further development.

2. System Performance. We evaluated the performance of our VC investment screening system. We found that our system's performance to be robust over time and better or comparable to previous results from the literature. However, we identified that our system does not adequately capture nuances of investment success. We outline ways to improve our system's performance in the future.

3. Model Evaluation. We developed a novel conceptual framework for VC investment decisions which guided our data source selection and feature creation. In evaluating our VC investment screening system, we also evaluated our conceptual framework. In that sense, we contribute a comprehensive, empirical study of startup performance. We discuss ways to build on our conceptual framework in future research.

5.1 System Design

The design of our automated Venture Capital (VC) investment screening system provides contributions to the literature including the use of data sources

CrunchBase and PatentsView, and the development of an adaptive pipeline optimisation system. In the following section, we discuss these contributions and areas for future development.

5.1.1 Data Collection & Preparation

Our system collects data from CrunchBase and PatentsView. Neither data source has been studied comprehensively in the literature. CrunchBase gained critical mass in 2012 and PatentsView was formed in 2015. CrunchBase and PatentsView offer a significant improvement regarding size and variety of features over previous data sources used in entrepreneurship research (e.g. surveys, interviews, closed datasets, etc.). However, we did find that collecting data from CrunchBase and PatentsView raised some issues that we needed to overcome.

We found the CrunchBase database to be highly sparse with many long-tailed features and irrelevant companies. These qualities have been identified in the literature previously [30] and are likely attributable to the crowd-sourced nature of CrunchBase. To address CrunchBase's dataset issues, we developed a classification pipeline with multiple pre-processing steps (as described in the next section). In addition, there is room for improvement in the implementation of our data collection system from CrunchBase. The current system downloads CSV dumps exported from the CrunchBase database. While this is an improvement on the status quo (manual data collection), a further improvement would be the development of a real-time connector to CrunchBase's API. CrunchBase publishes a daily changelog which our connector could use to request only information from companies that had changed. An entirely time-stamped database produced in this manner would also allow for greater analysis of temporal trends, and avoid the biases of our database slicing technique.

The primary issue we encountered during collection of PatentsView data was matching companies between PatentsView and CrunchBase. Companies often use variations on their names on these sources. We attempted to resolve this issue by standardising the names (e.g. removing suffixes and punctuation) and using Levenshtein distances to determine the likelihood that we were accurately matching the companies. We did observe a positive correlation between the number of patents filed by companies and their developmental stages which may suggest the matching was successful. However, PatentsView features did not contribute significantly to the models generated by our system despite previous studies that show patent filings to be key predictors of startup performance [6, 1]. Another issue that we encountered with PatentsView was that their API, unlike CrunchBase, does not provide a change-log. This limitation means our system

has to run a full database sweep (approximately 10 hours) to check its patent filing records are up-to-date. We hope PatentsView will add this functionality soon.

A key contribution of this project is our ability to capture the trajectory of startups and reliability of VC investment predictions through time. This contribution is only possible because of our database slicing technique, but this technique also raises concerns. We evaluated the slicing system leveraging both 'last-updated' and 'created-at' timestamps: using 'last-updated' timestamps excluded many recently-updated companies, whereas using created-at timestamps retained more features than is historically accurate. We decided to use created-at timestamps. While the relational database structure mediates this effect (e.g. acquisition and investment records have separate timestamps), this technique may inflate our system's performance on companies that have had many recent edits to their records. We attempted to evaluate this impact by comparing a CrunchBase database collected in December 2013 with a slice from our primary database collected in September 2016. We found minimal variance in the number of records found in each relation. However, because the database schema changed between 2013 and 2016, we were unable to determine whether the completeness of each respective record is similar. As we collect more database dumps, we will be able to evaluate this technique better.

5.1.2 Pipeline Optimisation

Previous studies that have applied data mining techniques to startup performance and VC investment have typically evaluated a few specific classification algorithms [11, 16, 31, 32]. In a novel contribution, we presented an adaptive pipeline optimisation process that provides greater accuracy and re-calibration of the system as the data sources change over time.

Pipeline creation performs a broad search and evaluation of candidate pipelines with varying hyper-parameters. This search is performed across the pre-processing steps of the pipeline and also the classification algorithms. We found classification algorithm tuning had the greatest impact on performance during optimisation. It appears little optimisation of the pre-processing steps were needed. In aggregate, the performance of the classification pipeline was not improved by the pre-processing steps. Interactions between the pre-processing steps and the classification algorithms (e.g. Random Forests are more resilient to low orthogonality than Naive Bayes) may have reduced this aggregate effect. Nonetheless, optimisation of the pre-processing steps should still improve the overall robustness of the optimisation process as the dataset and prediction tasks change.

Our literature review suggested Random Forests would be the most successful classifier, followed by Artificial Neural Networks and Support Vector Machines. We found Random Forests and Logistic Regressions performed best and Artificial Neural Networks and Support Vector Machines underperformed. Random Forests may have outperformed the other algorithms due to its robustness to missing values and irrelevant features. Learning curves also revealed that, unlike most of the other classifiers, Random Forests was least likely to converge early, which suggests with larger training sets it should perform better. Logistic Regression. It was surprising that Support Vector Machines and Artificial Neural Networks underperformed the other algorithms. However, these algorithms are harder to tune accurately, so it may reflect that our search process was too limited. In production, we could perform the search process over longer iterations which would likely result in better performance from these algorithms.

The second step in our system's pipeline optimisation process is pipeline selection. In this component, we rank the candidate pipelines generated previously and evaluate the best pipelines over different dataset slices. This process ensures our final pipeline is robust in its performance over time. We do not observe significant variance in the pipelines on aggregate against the dataset slices, but there is variance in each pipeline's performance against the dataset slices. The former result suggests our pipelines produce models that are robust over time (which we reinforce in our later experimentation). The latter result justifies this step in our process. In our preliminary evaluation of this test, we selected the top ten candidate pipelines. Although there is still a strong positive correlation between the pipeline's initial ranking and their scores, we can see there are some individual deviations. Importantly, the top-ranked pipeline from the first stage has a lower median score than the second-ranked pipeline. These results suggest it is optimal to evaluate the top 3-5 candidate pipelines in this manner.

5.1.3 Automation & Efficiency

A key benefit of our proposed VC investment screening system is that it will reduce the amount of manual effort required before an investment decision. The implementation of the system described in this paper goes some ways to address this. Currently the system is semi-autonomous: it has little requirement for external input besides configuration of investment criteria (e.g. forecast window, developmental stage, etc.), but still runs on-demand, rather than continuously.

An improved implementation of the system would run in the background continuously, scheduling components of the system to run as needed to ensure the results are always optimised. The system currently takes a total of 46 hours to

complete. Most of the duration of the system is because of pipeline creation, which performs a large search across potential pipeline hyper-parameters. However, when placed into production, this component could infrequently be run (e.g. once per year) to ensure the pipelines remain optimally suited for the dataset. The next component of the system, selecting the most robust pipeline, could occur more frequently (e.g. once every month). The final component of the system, making predictions, could be evaluated whenever the system collects new data (e.g. once per day) because it only takes an hour to complete.

When we evaluated the learning curves of the pipeline selected by our system, we found our system's performance had converged for some target outcomes but for others (e.g. IPO, Acquisitions) it had yet to converge. This finding suggests our system could still benefit from a larger dataset. However, these results could also arise because our experimental configuration uses a pipeline that was optimised for predicting our base target outcome. We might find other classification pipelines yield better performance if the entire system runs for different target outcomes. A key advantage of our system is that, as we collect more data over time, our pipeline optimisation process will adapt to the nature of the dataset and select classifiers with less bias and more variance so that we may see Support Vector Machines or Artificial Neural Networks adopted by the system in the future.

5.2 System Performance

We evaluated the performance of our Venture Capital (VC) investment screening system across a range of datasets with different training dates, forecast windows, developmental stages, and target outcomes. In this section, we discuss the performance of our system across these domains, with a comparison to previous systems from literature. We also suggest some limitations of our experimental design.

5.2.1 Historical Datasets

Robustness over time is a critical attribute for our system, so VC firms can rely that models trained on historical datasets will be accurate into the future. We evaluated the robustness of our system's performance by training the system on datasets of different dates from 2012-2014. We evaluated the performance of the system using a variety of evaluation metrics. Across all evaluation metrics, we observed little variance in performance. We observed that performance variance

decreased as forecast windows became longer in duration. This trend may be because longer-term models use fundamental features that are less likely to vary over time. This trend may also be because there is greater variance in the dates of datasets with shorter forecast windows because our database slicing technique is restricted in how far back we can reverse-engineer data slices before there are not enough observations left to proceed.

5.2.2 Forecast Window

While predicting whether a startup is likely to exit in the future is useful, predicting the timing of that exit is also critical because the time taken to exit is inversely related to the rate of return that the investor receives. We evaluated our system's performance against forecast windows of 2–4 years and a variety of evaluation metrics. We observe a positive relationship between performance and length of forecast window. This trend suggests that it is harder to predict when a startup company will raise a funding round or exit than whether it will do it at all. Non-performance related factors (e.g. finding an investor with the right fit, requiring extra funding to enter a new market) may influence the timing of each activity. In the future, it would be interesting to explore forecast windows closer in duration to a typical VC investment horizon (3–8 years) [3]. As we discovered in our preliminary evaluation, the majority of companies that are going to raise funds at all will raise funds over a three-year period. However, exits seem to operate on a longer lead-time. It is unlikely that our ability to predict additional funding rounds would change over a longer forecast window but we would expect to see our performance at predicting exits continue to increase. We also observe that a relationship between the magnitude of performance improvement over longer forecast windows and how sensitive the evaluation metric is to both the imbalanced nature of the dataset and our bias towards the positive class. Accordingly, F1 Scores show the greatest performance improvement and Area under the Receiver Operating Characteristic (ROC) curve shows little variance. This finding justifies our decision to select F1 Scores as our primary evaluation metric.

5.2.3 Developmental Stage

VC investment decision-making is difficult. In our preliminary evaluation of our dataset, we discovered that even firms at late developmental stages (e.g. Series B, C) who have received multiple rounds of screening, only have about a 30% chance of raising additional funding or exiting. We investigated how the perfor-

mance of our system varies across target companies of different developmental stages ranging from Pre-Seed through Series D+. We found a positive trend between later developmental stage and performance. Later-stage companies tend to have more available features in our dataset which may explain this trend. Beckwith (2016) studied companies seeking equity crowd-funding (which maps to Pre-Seed in our classification system) and showed poor classification results for predicting whether a company would raise the equity crowd-funding round [11]. Beckwith's highest F1 Score was 0.33, which is in line with the results we gained for more difficult prediction tasks (e.g. predicting IPO over three years). Stone (2014) suggested that VC investment screening was simply not viable before Series A stage [14] whereas we observed performance up to an F1 Score of 0.58 for predicting funding rounds and exits by Seed stage companies. Bhat (2011) studied companies that had previously raised three VC rounds (Series C in our classification) and received comparable results to our system [16].

A discrepancy in the positive trend between developmental stage and performance is a slight decrease in our system's performance at Series D+. This finding may be because the model is primarily predicting exits at this developmental stage, rather than additional funding rounds, and exits are harder to predict. To investigate this discrepancy further, we split the datasets into their developmental stages and fit the model onto each of these sub-datasets individually. Pre-Seed companies make up most of our original dataset, and we see the smallest improvement for this stage. However, for Series D+ we see a large improvement, which suggests the features that predict Series D+ performance vary from earlier stages. Overall, this stage-specific fit method results in a performance improvement, despite each model having significantly fewer observations on which to train. This finding suggests that the underlying factors that influence startup investment performance for each developmental stage are significantly different.

5.2.4 Target Outcome

Ultimately, our system seeks to identify startup investment opportunities that will return their invested funds many times over within an investment horizon of a VC's fund (typically 3–8 years) [3]. However, our dataset has little information about valuations at funding rounds or during acquisitions because valuation is considered sensitive and confidential. Instead, we developed broader target outcomes as rough proxies for the underlying success of the investment. These outcomes include raising additional funds, being acquired, having an IPO and combinations thereof. We evaluated each of these outcomes separately to determine their effect on our system's performance. Our system is better at predicting more common events, so performs best at predicting additional funding rounds

and worst at predicting IPOs. The system's poor performance on IPOs may also be due to non-performance related factors that affect IPO timing, like financial market conditions. Surprisingly, our system is better at predicting whether a company will exit, than whether it will exit or gain additional funding in combination. Other factors interacting with target outcome may be causing this effect. For example, most companies that exit are in their later stages of development, while most companies that raised additional funding rounds are in earlier stages of their development. Our system may not be capable of learning to disentangle these factors when predicting a combined outcome.

While our target outcomes provide a proxy for investment success, some nuances are not captured by these outcomes. While most funding rounds are generally at higher valuations than the previous round, some funding rounds are not — these are termed 'down-rounds'. Likewise, although most acquisitions are at higher valuations, sometimes they are not, and are instead used to recruit many of the staff that worked at the startup — these are sometimes termed 'acqui-hires'. As our publicly-sourced dataset has little information about valuations at funding rounds or during acquisitions, our system has little ability to distinguish between successful activity and down-rounds or acqui-hires. These discrepancies limit the performance of our system. Appendix G presents four case studies that highlight the nuances of our system's performance. In future, the application of sentiment analysis to media coverage of funding rounds, acquisitions or IPOs may indicate whether the raise or exit was genuinely successful.

5.2.5 Limitations

Our experimental design involved evaluating the performance of the system across a range of variables, including the size of each training set, date of each training set, duration of forecast window, company developmental stage, and target outcome. For each of these experiments, we manipulated these variables during the model fit and prediction step of our system design. However, to reduce the time taken by our experiments, we used the same optimised pipeline for each experiment (for the configuration, see Appendix E). This pipeline optimisation step takes the vast majority of time of our system (84.8%). By using a pipeline optimised for different objectives we are likely to have under-reported the performance of our system. In future research, it would be interesting to determine the extent to which the results of our pipeline optimisation changes with respect to these variables and the extent to which our results improve.

5.3 Model Evaluation

As a by-product of the evaluation of our system, we were also able to provide a comprehensive study of the determinants of startup investment performance. From our literature review, we developed a conceptual framework for startup investment performance, based on previous work by Ahlers and colleagues [4]. Our conceptual framework proposed that Venture Capital (VC) investment decisions have two primary components: startup potential and investment confidence. We decomposed these components into 15 factors identified in previous empirical studies in the literature. Our system evaluates many of these factors. Through our experimentation on the system, our system generated models that describe features associated with startup investment performance over time, with different forecast windows, developmental stages, and target outcomes.

5.3.1 Historical Datasets & Forecast Window

Within the VC industry, there is a commonly-held belief that startup performance is too volatile and qualitative to be predicted with any real accuracy using data mining techniques [14]. The models generated by our system did not provide evidence to support this assertion. We evaluated our system by training it on a range of historical datasets from 2012-16 across varying forecast windows. We found that models generated by the system were robust to changes in training date, with a standard deviation of less than 1% of the total normalised feature weights. We found a similar trend with respect to forecast window, with little variance for periods of 2–4 years. Despite this, we previously observed a positive relationship between system performance and forecast window length. Together, these findings suggest that for a given set of independent variables (developmental stage and target outcome), our models of startup investment performance are stable over time. This finding is in contradiction to the widely held within the VC industry that the factors that influence startup investment performance change over time. Admittedly, our models do not perfectly predict startup investment success, and dynamic factors not incorporated in our system may cause this margin of error. If not to a decision-making degree of performance, our results still suggest features that correlate with startup investment success are predictable and stable.

5.3.2 Developmental Stage

Previous studies have largely neglected to investigate how models of startup performance change across the startup development life-cycle. Most research in this field has studied either early-stage companies [11, 14, 13, 4], or much later-stage companies [16]. We believe that this is the first comprehensive study that takes a broader approach. We found considerable variation in models developed for companies of different stages. We find that *Advisors* is a more important factor to earlier stage companies than late stage companies. One explanation for this finding that successful startups recruit experienced influential advisors earlier to fill gaps in their experience but by the time these startups reach later rounds, this advantage is lessened by the influence of investors. We found that *Investors* becomes more important as companies approach an exit (Series D+). This finding may be due to influential investors having more experience at going through exit processes, and leveraging their contacts to make the process easier. We found that *Executives*, and *Founders* factors are important in Pre-Seed companies, which has previously been substantiated in the literature [11, 8]. This finding may be because at early stages of a company's development there is little to rely on aside from the previous experience and skill-set of the founding team and staff. Finally, we found that the *Economy* factor was most important at the Pre-Seed stage and has little effect at other stages. We did not expect economic factors to have a large effect on startup performance because, in comparison to larger, more established companies, startups are flexible, agile and have more focused markets. One explanation for the greater impact of the *Economy* factor at the Pre-Seed stage is that perhaps more people leave established companies to launch startups when the economy is doing poorly, but these companies do not survive longer than the Pre-Seed stage.

5.3.3 Target Outcome

The startup landscape is constantly changing and so is what is considered a successful VC investment. Amazon went public in 1997, just two years after its Series A, at a market capitalisation of $440M. Contrast that with Uber, which remains private six years on and recently raised $3.5B at a $59B pre-money valuation. While previous studies separated funding raises [11, 13, 4, 8] from predicting exits [16], we decided to evaluate our system across a variety of target outcomes that reflects this changing landscape. There is considerable variance between the models generated by our system to predict additional funding rounds, acquisitions and IPOs. The greatest predictor of additional funding rounds was *Advisors*, followed by *Executives*. The greatest predictor of acquisition was *Ex-*

ecutives, followed by *Investors* and *Founders*. The greatest predictor of Initial Public Offering (IPO) was *Investors*, followed by *Executives* and *Funding*. There is substantial evidence to support the finding that advisors, founders and executives predict startup performance [1, 5, 11]. Although a distinct model predicts each target outcome, all models are weighted towards human capital features. The high proportion of early stage companies in our dataset may inflate this effect because we would expect early stage companies to rely more heavily on human capital as they have few other resources. Aside from human capital, *Investors*, which covers the reputation and track record of invested VC firms, was a good predictor of acquisitions and IPOs. Investors who have a strong track record probably have had more experience helping portfolio companies exit and have more connections to leverage in this process. Most previous studies of startup funding focus on early-stage companies whereas studies that focus on exits tend to investigate later-stage companies. Our study bridges this divide in a way that is more holistic and more useful for investors.

5.3.4 Limitations

5.3.4.1 Missing Features

While CrunchBase and PatentsView provide features that cover much of our conceptual framework, there are factors we were unable to evaluate in this implementation of our system. Missing factors include media coverage, strategic alliances and financial performance. While it would likely be a significant factor in our models, we do not expect to be able to source financial information for our dataset in the future. In fact, we consider it to be a key benefit of our system that it can perform accurately without detailed financial information. The paucity of available financial data is what makes VC investment screening distinct from other fields of finance. Collecting data to support the other missing factors may be easier to source in the future. CrunchBase API provides an archive of media coverage on each startup company, so connecting directly to the CrunchBase API (rather than the CSV-dumps) would give access to this feature. Social influence is harder to capture because historical records of social media activity are hard to find. CrunchBase tracks whether companies have social media profiles but does not provide time-stamps for this information so we cannot use the information to create historical records. There are some Twitter historical data services, but these are expensive to use. Finally, strategic alliance information (e.g. with suppliers or universities) is not a typical feature that is recorded but could be engineered through textual analysis on media coverage. We should investigate these features in future work.

5.3.4.2 Simple Features

While our features covered a broad conceptual framework, we derive many of features from simple models (e.g. a company's location, the age of a company, the amount of funding a company has raised). While these types of features are consistent with most other studies in this field, the simplicity of these features may have reduced our system's ability to represent more complex factors. There is preliminary research that features derived from more complex models like semantic text features (e.g. keyword analysis from patents) [6, 13] and social network features (e.g. networks of social influence) [9, 7, 12] are significant predictors of startup performance. The factors that could benefit most from these features (e.g. patents and social influence) were under-represented in the models generated by our system, which supports this line of reasoning. There are disadvantages, however, in adopting more complex, dynamically generated features. A key contribution of this project was our ability to test our system's robustness by training on different datasets and then comparing the results. This process would not be possible if we did not maintain a consistent feature vector — we would be unable to train and test on different datasets because the features would not align.

Finally, while our project did incorporate longitudinal analyses (with respect to evaluating our system's performance against historical datasets and for different forecast windows), we did not specifically analyse temporal patterns in the trajectory of startups. Temporal analyses are an interesting area for future research. For example, probabilistic networks (e.g. Markov networks) could be used to represent the sequence and timing of different startup activities (e.g. media coverage, funding rounds, IPOs, hiring, etc.). Machine learning techniques could be applied to these networks to learn patterns of activity that are likely to lead to investment success. This technique has the potential to provide finer predictions than our cross-sectional models. It was probably not viable to apply these types of techniques to private companies in the past because not enough data was available. However, this project shows that CrunchBase and PatentsView cover a considerable feature set, particularly for later stage startups, and so this area deserves further investigation.

CHAPTER 6

Conclusions

Our project's aim was to produce an investment screening system suitable for use in the Venture Capital (VC) industry. Our system identifies startups that are likely to raise additional funding, become acquired or have an Initial Public Offering (IPO) (or some combination thereof) in a given period. While this is a challenging task, our system achieved results that have practical application for VC firms.

6.1 Evaluation of Criteria

We evaluated our system against three criteria: practicality, robustness, and versatility.

6.1.1 Practicality

Investment screening involves considerable time and effort for Venture Capital (VC) firms [33]. We evaluated whether it is practical for our system to replace existing screening processes (e.g. Google search, industry papers, databases). Our automated system is more efficient than these methods because it requires minimal user input. Our system takes 46 hours to run, which is reasonable because screening is not time-sensitive in this industry. The majority of the time taken by our system is due to pipeline optimisation. In the future, we could develop a scheduling system that runs time-intensive components (like pipeline optimisation) less frequently with minimal reduction in performance.

6.1.2 Robustness

VC firms are concerned that investment models trained on historical data will not accurately predict future trends and activity [14]. We evaluated whether our

system's predictions are robust over time. We found that training our system on different historical datasets had a minimal effect on the system's performance and the models it generated. This finding suggests that VC firms should be able to act on predictions made by our system. Our system is also robust to dataset size because it chooses an optimal classification pipeline based on the dataset available. It is likely that our system will continue to improve in performance as its data sources grow.

6.1.3 Versatility

VC firms vary in the investments they make according to their interests, the lifecycles of their funds, and the portfolios that they hold [3]. A VC investment screening system should be versatile in its ability to work across these variables. We evaluated our system's ability to perform across a large domain of investment prediction tasks. Tasks included predicting different target outcomes (e.g. Initial Public Offering (IPO), acquisition) for companies at different developmental stages (e.g. Seed, Series A, etc.) over different forecast windows (e.g. 2–4 years). These variables have significant effects on our system's performance and the models it generates. Where comparable, our system produced better or similar results to previous attempts.

6.2 Future Work

We identified several areas of further investigation that build on our work.

6.2.1 Systems Integration

Our system was designed to meet criteria critical to Venture Capital (VC) firms, but we must now assess how to integrate our software into their systems. First, we hope to conduct a use case study with one or more VC firms. This study will inform the commercialisation of our system. One potential area of further development is an autonomous system that schedules the system components to ensure consistent near-optimal performance. This task scheduling system would pair well with an improved CrunchBase data collection system that connects to the CrunchBase API rather than downloading CSV-dumps. An API connector would allow the task scheduler to run an optimisation process that maximises performance improvement against dataset changes and time taken.

6.2.2 Feature Improvement

This project provided a comprehensive study of features that predict startup performance, but there remain improvements we can make to the feature set. Some factors from our framework were not represented or supported by few features: media coverage, strategic alliances, financial performance, social influence, and industry performance. We expect their inclusion to improve our predictions. We also expect that more complex features like semantic text analysis (e.g. keyword analysis from patents, sentiment analysis from media) and social network analysis (e.g. co-investment networks, spheres of influence) will improve our predictions. Finally, we hope to explore whether temporal relationships between startup activities (e.g. media coverage, funding rounds, Initial Public Offering (IPO), etc.) might also improve performance.

6.3 Summary

We set out to create a Venture Capital (VC) investment screening system that met criteria of practicality, robustness, and versatility, and we have indeed created such a system. The work required to achieve this project's goals was extensive, from reviewing the state of VC theory and data mining, to developing systems that collect data from CrunchBase and PatentsView, to developing an adaptive classification pipeline process, and finally performing experiments that validate the ability of the system to meet our criteria above.

This project makes three primary contributions with implications for industry and research. First, we designed our system for the VC industry: it is near-autonomous, robust to changes in dataset and prediction task, and uses a comprehensive feature set collected from large public online databases. Second, our system's performance is not only better or comparable to previous studies, but it also addresses a far larger domain of investment prediction tasks with respect to forecast window, developmental stage and target outcome. Third, this project contributes an empirical study of models of startup investment performance more comprehensive than any found in the literature. Ultimately, this project makes steps towards automation in the VC industry.

APPENDIX A

Data Sources

This appendix provides an extended overview of potential online data sources relevant to developing a Venture Capital (VC) investment screening system, including startup databases, social networks, and other sources.

A.1 Startup Databases

Databases play a critical role in understanding the startup ecosystem, aggregating information about startups, investors, media and trends. Most startup databases are closed systems that require commercial licenses (e.g. CB Insights, ThomsonOne, Mattermark). CrunchBase and AngelList are two crowd-sourced and free-to-use alternatives.

A.1.1 CrunchBase

CrunchBase is an open online database of information about startups, investors, media coverage and trends, focusing on high-tech industry in the United States. It relies on its active online community to contribute to and edit most of its pages. However, this results in unpopular startups having relatively sparse profiles. CrunchBase has three provisions to prevent and remediate inaccurate crowd-sourced entries. First, users authenticate their accounts with a social media account which allows CrunchBase to verify a user's identity. Second, every change goes through a machine review, which flags significant or questionable updates. Third, established startups have their editing privileges locked and updates require manual verification.

A.1.2 AngelList

AngelList combines the functionality of an equity crowdfunding platform, a social networking site and an online startup database. As an equity crowdfunding platform, users create profiles for their startups on AngelList, and use the platform to attract investment. Investors use the platform to identify investment opportunities and can invest directly through AngelList, often alongside other investors in investment syndicates. AngelList is also an online startup database. It has a data-sharing agreement with CrunchBase which results in significant overlap between the two sources, though CrunchBase tends to have more comprehensive records of funding rounds [12]. AngelList tracks "startup roles" (e.g. founders, investors, employees) with a creation-time, start-time and end-time. This means that, unlike CrunchBase, AngelList's networks can be re-created through time, which is useful for longitudinal studies.

A.1.3 Comparison

AngelLists primary function is as an equity crowdfunding platform but it has a data-sharing agreement with CrunchBase which results in significant overlap between the two sources. CrunchBase tends to have more comprehensive records of funding rounds [12] and media coverage but AngelList also has a social network element where users can 'follow each other – in a similar way to Twitter. The crowd-sourced nature of CrunchBase and AngelList has advantages and limitations. The key advantages are that access to the databases is free and the dataset is relatively comprehensive. The limitations are that both CrunchBase and AngelList have relatively sparse profiles (i.e. limited depth), particularly for unpopular startups.

A.2 Social Networks

Social networks provide an interesting perspective into the process of opportunity discovery and capitalisation that characterises entrepreneurship. Two social networks studied in detail in entrepreneurship research are LinkedIn and Twitter.

A.2.1 LinkedIn

LinkedIn is a massive professional social network often used in studies of entrepreneurship for measures of employment, education and weak social links.

These measures are difficult to collect elsewhere. In addition, LinkedIn can provide a measure of the professional influence of founders and investors. Unfortunately, as of May 2015, the LinkedIn API no longer allows access to authenticated users' connection data or company data, making it difficult to use for social network analyses.

A.2.2 Twitter

Twitter is a massive social networking and micro-blogging service which is studied in entrepreneurship research because it is used by founders, investors, and customers to quickly communicate and broadcast. Twitter is a directed network where users can follow other users without gaining their permission to do so. Twitter's public API provides access to social network topological features (e.g. who follows who) and basic profile information (e.g. user-provided descriptions). However, Twitter's API only provides Tweets published within the last 7 days and access to historical Twitter data requires a commercial license.

A.3 Other Sources

While startup databases and social networks provide a variety of information on startups, there are two important areas that they do not cover: patent filings and financial performance.

A.3.1 PatentsView

Startups often file patents to apply for a legal right to exclude others from using their inventions. In 2015, the US Patents Office (USPTO) launched PatentsView, a free public API to allow programmatic access to their database. PatentsView holds over 12 million patent filings from 1976 onwards. The database provides comprehensive information on patents, their inventors, their organisations, and locations. It may be difficult to match identities across PatentsView to other data sources because registered company names (as in PatentsView) are not always the same as trading names (as elsewhere).

A.3.2 PrivCo

Finding other information on startups, like financial information, is difficult. Unlike public companies, private companies are not required to file with the United

States Securities and Exchange Commission (or international equivalent). Proprietary databases provide some data on private companies but commercial licenses are prohibitively expensive and have poor coverage of early-stage companies. PrivCo is one of few commercial data sources for private company business and financial intelligence. PrivCo focuses its coverage on US private companies with at least $50-100 million in annual revenues but also has some coverage on smaller but high-value private companies (like startups).

APPENDIX B

Classification Algorithms

This appendix provides an extended overview of potential classification algorithms relevant to developing a Venture Capital (VC) investment screening system. The characteristics of each algorithm is described and then applications of the algorithm to problems adjacent to VC investment are reviewed.

B.1 Naive Bayes

Naive Bayes is a simple generative learning algorithm. It is a Bayesian Network that models features by generating a directed acyclic graph, with the strong (naive) assumption that all features are independent. While this assumption is generally not true, it simplifies estimation which makes Naive Bayes more computationally efficient than other learning algorithms. Naive Bayes can be a good choice for data sets with high dimensionality and sparsity as it estimates features independently. Naive Bayes sometimes outperforms more complex machine learning algorithms because it is reasonably robust to violations of feature independence [23]. However, Naive Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features [28]. Naive Bayes was used alongside Logistic Regression, Decision Trees and Support Vector Machines to predict success in equity crowdfunding campaigns on the AngelList data set [11]. None of these models performed well. The algorithm that best predicts startup investment was Naive Bayes with a Precision of .41 and Recall of .19, which means only 19% of funded startups were classified correctly by the model. The author suggests the poor performance of their algorithms is caused by features not captured in their data set relating to Intellectual Capital, Third Party Validation and Historical Performance. These features will be included in this study.

B.2 Logistic Regression

Regression is a class of statistical methods that investigates the relationship between a dependent variable and a set of independent variables. Logistic regression is regression where the dependent variable is discrete. Like linear regression, logistic regression optimises an equation that multiplies each input by a coefficient, sums them up, and adds a constant. However, before this optimisation takes place the dependent variable is transformed by the log of the odds ratio for each observation, creating a real continuous dependent variable on a logistic distribution. A strength of Logistic Regression is that it is trivial to adjust classification thresholds depending on the problem (e.g. in spam detection [28], where specificity is desirable). It is also simple to update a Logistic Regression model using online gradient descent, when additional training data needs to be quickly incorporated into the model (incremental learning). Logistic Regression tends to underperform against complex algorithms like Random Forest, Support Vector Machines and Artificial Neural Networks in higher dimensions [21]. This underperformance is observed when Logistic Regression is applied to startup investment prediction tasks [11, 16]. However, weaker predictive performance has not prevented Logistic Regression from being commonly used. Its simplicity and ease-of-use means it is often used without justification or evaluation [5].

B.3 K-Nearest Neighbours

K-Nearest Neighbours is a common lazy learning algorithm. Lazy learning algorithms do not produce explicit general models, but compare new instances with instances from training stored in memory. K-Nearest Neighbours is based on the principle that the instances within a data set will exist near other instances that have similar characteristics. K-Nearest Neighbours models depend on how the user defines distance between samples; Euclidean distance is a commonly used metric. K-Nearest Neighbour models are stable compared to other learning algorithms and suited to online learning because they can add a new instance or remove an old instance without re-calculating [23]. A shortcoming of K-Nearest Neighbour models is that they can be sensitive to the local structure of the data and they also have large in-memory storage requirements. K-Nearest Neighbours was compared to Artificial Neural Networks to predict firm bankruptcy [31]. K-Nearest Neighbours is attractive in bankruptcy prediction because it can be updated in real-time. By optimising feature weighting and instance selection, the authors improved the K-Nearest Neighbours algorithm to the extent that it outperformed the Artificial Neural Networks.

B.4 Decision Trees

Decision Trees use recursive partitioning algorithms to classify instances. Each node in a Decision Tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Methods for finding the features that best divide the training data include Information Gain and Gini Index [23]. Decision Trees are close to an "off-the-shelf" learning algorithm. They require little pre-processing and tuning, are interpretable to laypeople, are quick, handle feature interactions and are non-parametric. However, Decision Trees are prone to overfitting and have poor predictive power [34]. These shortcomings are addressed with pruning mechanisms and ensemble methods like Random Forests, respectively. Decision Trees were compared with Naive Bayes and Support Vector Machines to predict investor-startup funding pairs using CrunchBase social network data [32]. Decision Trees had the highest accuracy and are desirable because their reasoning is easily communicated to startups.

B.5 Random Forests

Random Forests are an ensemble learning technique that constructs multiple Decision Trees from bootstrapped samples of the training data, using random feature selection [35]. Prediction is made by aggregating the predictions of the ensemble. The rationale is that while each Decision Tree in a Random Forest may be biased, when aggregated they produce a model robust against over-fitting. Random Forests exhibit a performance improvement over a single Decision Tree classifier and are among the most accurate learning algorithms [34]. However, Random Forests are more complex than Decision Trees, taking longer to create predictions and producing less interpretable output. Random Forests were used to predict private company exits using quantitative data from ThomsonOne [16]. Random Forests outperformed Logistic Regression, Support Vector Machines and Artificial Neural Networks. This may be because the data set was highly sparse, and Random Forests are known to perform well on sparse data sets [35].

B.6 Support Vector Machines

Support Vector Machines are a family of classifiers that seek to produce a hyperplane that gives the largest minimum distance (margin) between classes. The key to the effectiveness of Support Vector Machines are kernel functions. Kernel functions transform the training data to a high-dimensional space to improve its

resemblance to a linearly separable set of data. Support Vector Machines are attractive for many reasons. They have high predictive power [34], theoretical limitations on overfitting, and with an appropriate kernel they work well even when data is not linearly separable in the base feature space. Support Vector Machines are computationally intensive and complicated to tune effectively (compared to Random Forests, for example). Support Vector Machines were compared with back-propagated Artificial Neural Networks in predicting the bankruptcy of firms using data provided by Korea Credit Guarantee Fund [36]. Support Vector Machines outperformed Artificial Neural Networks, possibly because of the small data set.

B.7 Artificial Neural Networks

Artificial Neural Networks are a computational approach based on a network of neural units (neurons) that loosely models the way the brain solves problems. An Artificial Neural Network is broadly defined by three parameters: the interconnection pattern between the different layers of neurons, the learning process for updating the weights of the interconnections, and the activation function that converts a neuron's weighted input to its output activation. A supervised learning process typically involves gradient descent with back-propagation [28]. Gradient descent is an optimisation algorithm that updates the weights of the interconnections between the neurons with respect to the derivative of the cost function (the weighted difference between the desired output and the current output). Back-propagation is the technique used to determine what the gradient of the cost function is for the given weights, using the chain rule. Artificial Neural networks tend to be highly accurate but are slow to train and require significantly more training data than other machine learning algorithms. Artificial Neural Networks are also a black box model so it is difficult to reason about their output in a way that can be effectively communicated. Artificial Neural Networks are rarely applied to startup investment or performance prediction because research in this area typically uses small and low-dimensional data sets. As one author puts it "More complex classification algorithms - artificial neural networks, Restricted Bolzmann machines, for instance - could be tried on the data set, but marginal improvements would likely result." [11]. However, this study will address these issues so Artificial Neural Networks may be more competitive.

APPENDIX C

Database Schema

This appendix provides the database schema for our system's master relational database. This database is primarily composed of data collected from CrunchBase with the exception of the Patents relation which is collected from PatentsView.

Relation	Attributes
Acquisitions	Acquiree Name, Acquiree Country, Acquiree State, Acquiree Region, Acquiree City, Acquirer Name, Acquirer Country, Acquirer State, Acquirer Region, Acquirer City, Acquisition Date, Acquisition Price, Acquisition Price Currency, Acquisition Price (USD), Acquiree CB URL, Acquirer CB URL, Acquiree UUID, Acquirer UUID, Acquisition UUID, Created Timestamp, Updated Timestamp
Category Groups	Category Group UUID, Category Name, Category Group List
Competitors	Entity UUID, Competitor UUID, Created Timestamp, Updated Timestamp
Customers	Entity UUID, Customer UUID, Created Timestamp, Updated Timestamp
Event Relationships	Event UUID, Entity UUID, Event Type, Relationship to Event (Type), Relationship to Event (Detail), Created Timestamp, Updated Timestamp
Events	Event UUID, Event Name, Short Description, Started Date, Ended Date, Registration Details, Registration URL, Start Time, End Time, Venue Name, Venue Address, Location UUID, Cost, Description, City, Region, Country, Continent, Permalink, CB URL, Logo URL, Profile Image URL, Event Roles, Created Timestamp, Updated Timestamp
Funding Rounds	Company Name, Country, State, Region, City, Company Category List, Funding Round Type, Funding Round Code, Announced Date, Raised Amount, Raised Amount Currency, Raised Amount (USD), Target Money Raised, Target Money Raised Currency, Target Money Raised (USD), Post Money Valuation, Post Money Valuation Currency, Post Money Valuation (USD), Investor Count, Investor Names, CB URL, Company UUID, Funding Round UUID, Created Timestamp, Updated Timestamp
Funds	Entity UUID, Fund UUID, Fund Name, Started Date, Announced Date, Raised Amount, Raised Amount Currency, Created Timestamp, Updated Timestamp
Investment Partners	Funding Round UUID, Investor UUID, Partner UUID
Investments	Funding Round UUID, Investor UUID, Is Lead Investor
Investors	Investor Name, Primary Role, Website Domain, Country, State, Region, City, Investor Type, Investment Count, Total Funding (USD), Founded Date, Closed Date, CB URL, Logo URL, Profile Image URL, Twitter URL, Facebook URL, Investor UUID, Updated Timestamp
IPOs	Company Name, Country, State, Region, City, Stock Exchange Symbol, Stock Symbol, IPO Date, Opening Share Price, Opening Share Price Currency, Opening Share Price (USD), CB URL, IPO UUID, Company UUID, IPO UUID, Created Timestamp, Updated Timestamp
Jobs	Person UUID, Organization UUID, Started Date, Ended Date, Is Current, Job Title, Job Role, Is Executive Role, Is Advisory Role
Organizational Parents	Organization UUID, Parent Organization UUID, Relationship to Parent, Created Timestamp, Updated Timestamp
Organization Descriptions	Organization UUID, Description
Organizations	Company Name, Primary Role, Website Domain, Website URL, Country, State, Region, City, Zipcode, Address, Operating Status, Short Description, Category List, Category Group List, Number of Funding Rounds, Funding Total (USD), Founded Date, First Funding Date, Last Funding Date, Closing Date, Employee Count, Company Email, Company Phone, Facebook URL, CB URL, Logo URL, Profile Image URL, Twitter URL, Organization UUID, Created Timestamp, Updated Timestamp
Patents	Assignee UUID, Patent Date, Citations by Patents, Citations of Patents, Patent Type
People	First Name, Last Name, Country, State, City, CB URL, Logo URL, Profile Image URL, Twitter URL, Facebook URL, Primary Affiliation Organization, Primay Affiliation Title, Primary Organization UUID, Gender, People UUID, Created Timestamp, Updated Timestamp
People Descriptions	People UUID, Description

Table C.1: Relational database schema.

APPENDIX D

Pipeline Hyperparameters

This appendix describes the search space of our system's pipeline optimisation process. The pipeline consists of an imputer, transformer, scaler, extractor and classifier in sequence. Each function has a number of hyperparameters that are either constant, iterables or distributions to be sampled.

Imputer	[Mean, Median, Most Frequent]
Transformer	[None, numpy.log1p, numpy.sqrt]
Scaler	
None	
StandardScaler	With Mean: True, With STD: True
RobustScaler	With Centering: True, With Scaling: True, Quantile Range: (25, 75)
MinMaxScaler	Feature Range: (0, 1)
Extractor	Function: PCA, Components: In Range (1, 100), Whiten: False, SVD Solver: Auto
Classifier	
Naive Bayes	
K-Nearest Neighbours	Neighbors: In Range (5,20), Weights: [Uniform, Distance], Algorithm: Auto, Leaf Size: 30, Metric: Minkowski, Distance: Euclidean
Logistic Regression	C: In Range (1e-3, 1e6), Penalty: [L1, L2], Solver: Liblinear, Fit Intercept: True, Intercept Scaling: True, Class Weight: Balanced, Tolerance: 1e-4
DecisionTree	Max Depth: In Range (5, 20), Criterion: [Gini, Entropy], Class Weight: Balanced, Splitter: Best, Max Features: None, Min Samples Split: 2, Min Samples Leaf: 1, Min Impurity Split: 1e-7
RandomForest	Estimators: In Range (10, 100), Max Depth: In Range (5, 20), Criterion: [Gini, Entropy], Class Weight: Balanced, Max Features: SQRT(Features), Min Samples Split: 2,Min Samples Leaf: 1,Min Impurity Split: 1e-7, Bootstrap: True
Support Vector Machine	C: [1e-5, 1e6], Probability: True, Class Weight: Balanced, Tolerance: 1e-3, Kernel: Linear, Poly: Degree: 3, Gamma: 1/Features, Coef0: 0, RBF: Gamma: 1/Features, Sigmoid: Gamma: 1/Features, Coef0: 0
Artificial Neural Network	Hidden Layers: 1, Hidden Layer Size: 100, Activation Function: [Identity, Logistic, Tanh, Relu], Alpha: [1e-3, 1e6], Solver: Adam Beta1: 0.9, Beta2: 0.999, Epsilon: 1e-8, Batch Size: min(200, Samples), Max Iterations: 200, Tolerance: 1e-4, Initial Learning Rate: 1e-3

Table D.1: Pipeline hyper-parameter search space.

APPENDIX E

Experimental Configuration

This appendix outlines the configuration of our system for the experiments outlined in Chapter 4. The base configuration is tested across nine database slices: three variations for each of three forecast windows. For the purposes of our experimentation, our optimised pipeline is held stable but in practice it is likely that the pipeline optimisation would be re-run for different configurations.

Base Configuration	
Database Slices	9 (3 x Forecast Window)
Forecast Window	3 [2, 3, 4]
Classification Pipeline	
Imputer	Most Frequent
Transformer	SQRT
Scaler	MinMaxScaler
Extractor	None, PCA
Classifier	Algorithm: Random Forest, Classes: Balanced, Criterion: Entropy, Bootstrap: True, Estimators: 34, Max Depth: 8, Max Features: SQRT(Features)
Experiment 1: Time Frame	
Developmental Stage	All (Combined)
Target Outcome	Extra Stage
Training Set Size	100%
Experiment 2: Developmental Stage	
Developmental Stage	All (Combined), All (Individual)
Target Outcome	Extra Stage
Training Set Size	100%
Experiment 3: Target Outcome	
Developmental Stage	All (Combined)
Target Outcome	Extra Stage, Extra Round, Exit, Acquisition, IPO
Training Set Size	100%
Experiment 4: Training Set Size	
Developmental Stage	All (Combined)
Target Outcome	Extra Stage
Training Set Size	1.0%, 3.2%, 10.0%, 31.7%, 100%

Table E.1: Experimental configuration.

APPENDIX F

Classification Reports

This appendix outlines the classification reports for our various experiments.

Slice Date		N	%	Accuracy	Precision	Recall	F1
2012	Positive	11,717	9.7	-	0.28	0.52	0.36
	Negative	109,312	90.3	-	0.94	0.86	0.90
	Avg/Total	121,029	-	0.82	0.88	0.82	0.85
2013	Positive	11,713	9.7	-	0.27	0.55	0.36
	Negative	109,117	90.3	-	0.95	0.84	0.89
	Avg/Total	120,830	-	0.81	0.88	0.81	0.84
2014	Positive	11,717	9.7	-	0.27	0.60	0.37
	Negative	109,312	90.3	-	0.95	0.82	0.88
	Avg/Total	121,029	-	0.80	0.88	0.80	0.83

Table F.1: Classification report by training set date. Forecast Window: 2 years. Target Outcome: Extra Stage. Developmental Stages: All. Training Set Size: Full.

Training Set Size		N	%	Accuracy	Precision	Recall	F1
485	Positive	10,133	18.0	-	0.47	0.20	0.28
	Negative	46,179	82.0	-	0.84	0.95	0.89
	Avg/Total	56,312	-	0.82	0.78	0.82	0.78
1,534	Positive	10,328	18.0	-	0.56	0.43	0.49
	Negative	47,033	82.0	-	0.88	0.93	0.90
	Avg/Total	57,361	-	0.84	0.82	0.84	0.83
4,851	Positive	10,939	18.0	-	0.55	0.54	0.55
	Negative	49,739	82.0	-	0.90	0.90	0.90
	Avg/Total	60,678	-	0.84	0.84	0.84	0.84
15,339	Positive	12,803	18.0	-	0.51	0.62	0.56
	Negative	58,363	82.0	-	0.91	0.87	0.89
	Avg/Total	71,166	-	0.82	0.84	0.82	0.83
48,506	Positive	18,642	18.0	-	0.49	0.64	0.56
	Negative	85,691	82.0	-	0.92	0.85	0.88
	Avg/Total	104,333	-	0.82	0.84	0.82	0.83

Table F.2: Classification report by training set size. Forecast Window: 4 years. Target Outcome: Extra Stage. Developmental Stages: All. Training Set Dates: 2012-2014 (3)

Forecast Window		N	%	Accuracy	Precision	Recal	F1
2 Years	Positive	35,147	9.7	-	0.27	0.56	0.36
	Negative	327,741	90.3	-	0.95	0.84	0.89
	Avg/Total	362,888	-	0.81	0.88	0.81	0.84
3 Years	Positive	37,169	14.3	-	0.40	0.62	0.48
	Negative	222,540	85.7	-	0.93	0.84	0.88
	Avg/Total	259,709	-	0.81	0.85	0.81	0.83
4 Years	Positive	30,279	18.0	-	0.49	0.62	0.55
	Negative	137,646	82.0	-	0.91	0.86	0.88
	Avg/Total	167,925	-	0.82	0.84	0.82	0.82

Table F.3: Classification report by forecast window. Target Outcome: Extra Stage. Developmental Stages: All. Training Set Size: Full. Training Set Dates: 2012-2014 (3)

Developmental Stage		N	%	Accuracy	Precision	Recall	F1
Pre-Seed	Positive	16,038	13.0	-	0.39	0.54	0.45
	Negative	107,760	87.0	-	0.93	0.88	0.90
	Avg/Total	123,798	-	0.83	0.86	0.83	0.84
Seed	Positive	5,472	21.7	-	0.55	0.62	0.58
	Negative	19,692	78.3	-	0.89	0.86	0.87
	Avg/Total	25,164	-	0.81	0.82	0.81	0.81
Series A	Positive	3,534	41.0	-	0.72	0.72	0.72
	Negative	5,082	59.0	-	0.81	0.80	0.80
	Avg/Total	8,616	-	0.77	0.77	0.77	0.77
Series B	Positive	2,682	48.9	-	0.81	0.66	0.73
	Negative	2,802	51.1	-	0.73	0.85	0.79
	Avg/Total	5,484	-	0.76	0.77	0.76	0.76
Series C	Positive	1,620	55.5	-	0.85	0.77	0.81
	Negative	1,299	44.5	-	0.74	0.83	0.78
	Avg/Total	2,919	-	0.80	0.80	0.80	0.80
Series D+	Positive	933	48.0	-	0.91	0.73	0.81
	Negative	1,011	52.0	-	0.79	0.93	0.86
	Avg/Total	1,944	-	0.84	0.85	0.84	0.83

Table F.4: Classification report by developmental stage. Forecast Window: 4 years. Target Outcome: Extra Stage. Training Set Dates: 2012-2014 (3). Training Set Size: Full

Target Outcome		N	%	Accuracy	Precision	Recall	F1
Extra Stage	Positive	30,279	18.0	-	0.50	0.61	0.55
	Negative	137,646	82.0	-	0.91	0.86	0.89
	Avg/Total	167,925	-	0.82	0.84	0.82	0.83
Extra Round	Positive	21,588	12.9	-	0.34	0.61	0.44
	Negative	146,337	87.1	-	0.93	0.83	0.88
	Avg/Total	167,925	-	0.80	0.86	0.80	0.82
Exit	Positive	12,372	17.4	-	0.46	0.70	0.56
	Negative	155,553	92.6	-	0.97	0.94	0.96
	Avg/Total	167,925	-	0.92	0.94	0.92	0.93
Acquisition	Positive	10,566	16.3	-	0.40	0.71	0.51
	Negative	157,359	93.7	-	0.98	0.93	0.95
	Avg/Total	167,925	-	0.91	0.94	0.91	0.93
IPO	Positive	2,052	1.2	-	0.28	0.57	0.37
	Negative	165,873	98.8	-	0.99	0.98	0.99
	Avg/Total	167,925	-	0.98	0.99	0.98	0.98

Table F.5: Classification report by target outcome. Forecast window: 4 years. Developmental Stages: All. Training Set Dates: 2012-2014 (3). Training Set Size: Full.

APPENDIX G

Case Studies

This appendix provides four case studies that highlight the nuances in the performance of our Venture Capital (VC) investment screening system. In particular, they highlight the limitations of our system with respect to understanding unsuccessful exits and funding rounds (e.g. down-rounds, acqui-hires).

	Company			
	ChaCha	Doctor.com	Fab	Mixpanel
Feature (2013-04-09)				
Age (Years)	7.4	0.6	4.3	3.8
Funding Raised ($m)	92.0	0.0	171.0	12.0
Funding Rounds (N)	8	0	8	4
Developmental Stage	Series D+	Pre-Seed	Series C	Series A
Predicted Outcome	✓	✗	✓	✓
Outcome (2017-04-04)				
Age (Years)	11.4	4.6	8.3	7.8
Funding Raised ($m)	96.0	5.0	336.0	77.0
Funding Rounds (N)	9	3	11	5
Developmental Stage	Series D+	Series A	Acquired	Series B
Actual Outcome	✗	✓	✓	✓
Correct Prediction	✗	✗	✓	✓

Table G.1: Company profiles and predictions.

1. ChaCha is an Indiana-based mobile Q&A service, launched in 2005. ChaCha has a long and convoluted investment history. It raised its Series A round in 2006 backed by Jeff Bezos of Amazon, before raising Series B-F rounds in 2007-10 to total funds of $92m. However, ChaCha took on additional rounds at lower valuations in 2011 and 2013. Our system predicted that ChaCha would raise funds or exit within the period of April 2013-2017. ChaCha did not take on any additional rounds and eventually closed in

2016. Our system did not predict this outcome accurately. As our publicly-sourced dataset has little information about valuations at funding rounds (valuation is considered more sensitive than quantum raised), our system has little ability to distinguish between succesful funding rounds and down-rounds (where valuation drops).

2. Doctor.com is a New York-based marketing automation platform for medical practices, launched in 2012. Doctor.com entered a three-year health-tech startup accelerator run by GE and StartUp Health in mid-2013. Our system did not predict that Doctor.com would raise funds or exit within the period of April 2013-2017. However, Doctor.com raised a $5m Series A round from Spring Mountain Capital in Feb 2017. This was a difficult prediction problem for our system. There was very little information about Doctor.com in 2013 and the Series A funding round came very late in the forecast window.

3. Fab is a New York-based e-commerce startup, launched in 2009. Fab raised $171m according to CrunchBase records, from reputable investors like Andreessen Horowitz, Mayfield Fund and First Round Capital, and once was reportedly valued at more than $1 billion. Our system predicted that Fab would raise funds or exit within the period of April 2013-2017. Later in 2013, Fab completed a Series D round for $150m. In 2015, Fab was acquired by PCH, reportedly for only a sum of ˜$20m. In this case, our system was technically accurate: Fab both raised funds and completed an exit. However, this exit was not a success for investors.

4. Mixpanel is a California-based consumer analytics platform, launched in 2009. Mixpanel came out of famed startup accelerator Y-Combinator and raised $12m from Seed - Series A rounds up to 2012, from reputable VC firms and angels like Sequoia Capital, Andreessen Horowitz and Max Levchin. Our system predicted that Mixpanel would raise funds or exit within the period of April 2013-2017. Mixpanel went on to raise a $65m Series B round from Andreessen Horowitz in December 2014 that valued the ccompany at $865m. However, during 2016, MixPanel cut 20 staff (primarily in sales) as it restructures towards higher profitability. This was a good prediction by our system over this forecast window but MixPanel's long term outlook is unclear.

Bibliography

[1] Baum, J. A., and Silverman, B. S. Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing* 19, 3 (2004), pp. 411–436.

[2] Graham, P. *Startup Investing Trends*. `http://www.paulgraham.com/invtrend.html/`. Online; accessed 15 May 2017. 2013.

[3] Gompers, P. A. Optimal investment, monitoring, and the staging of venture capital. *The Journal of Finance* 50, 5 (1995), pp. 1461–1489.

[4] Ahlers, G. K., et al. Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice* 39, 4 (2015), pp. 955–980.

[5] Gimmon, E., and Levie, J. Founder's human capital, external investment, and the survival of new high-technology ventures. *Research Policy* 39, 9 (2010), pp. 1214–1226.

[6] Hoenen, S., et al. The diminishing signaling value of patents between early rounds of venture capital financing. *Research Policy* 43, 6 (2014), pp. 956–989.

[7] Yu, Y., and Perotti, V. Startup Tribes: Social Network Ties that Support Success in New Firms. In: *Proceedings of 21st Americas Conference on Information Systems*. 2015.

[8] An, J., Jung, W., and Kim, H.-W. A Green Flag over Mobile Industry Start-Ups: Human Capital and Past Investors as Investment Signals. In: *PACIS 2015 Proceedings*. AIS Electronic Library, 2015, p. 67.

[9] Werth, J. C., and Boeert, P. Co-investment networks of business angels and the performance of their start-up investments. *International Journal of Entrepreneurial Venturing* 5, 3 (2013), pp. 240–256.

[10] Croce, A., Guerini, M., and Ughetto, E. Angel Financing and the Performance of High-Tech Start-Ups. *Journal of Small Business Management* (2016). issn: 1540-627X. doi: `10.1111/jsbm.12250`.

[11] Beckwith, J. Predicting Success in Equity Crowdfunding. Unpublished thesis. Joseph Wharton Research Scholars. Available at `http://repository.upenn.edu/joseph_wharton_scholars/25`. 2016.

[12] Cheng, M., et al. Collection, exploration and analysis of crowdfunding social networks. In: *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*. ACM. 2016, pp. 25–30.

[13] Yuan, H., Lau, R. Y., and Xu, W. The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems* 91 (2016), pp. 67–76.

[14] Stone, T. R. Computational analytics for venture finance. `http://discovery.ucl.ac.uk/1453383/1/Stone-TR-Computational-analytics-for-venture-finance.pdf`. PhD thesis. University College London, 2014.

[15] Patil, A. *CrunchBase's Venture Program Members Are Making Startup Data Better Than Ever*. `https://info.crunchbase.com/2015/01/crunchbases-venture-program-members-are-making-startup-data-better-than-ever/`. Online; accessed 18 05 2015. Jan. 2015.

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

[17] Fried, J. M., and Ganor, M. Agency costs of venture capitalist control in startups. *New York University Law Review* 81 (2006), p. 967.

[18] Shan, Z., Cao, H., and Lin, Q. Capital Crunch: Predicting Investments in Tech Companies. Unpublished thesis. Stanford University. Available at `http://www.zifeishan.org/files/capital-crunch.pdf`. 2014.

[19] Conti, A., Thursby, M., and Rothaermel, F. T. Show Me the Right Stuff: Signals for High-Tech Startups. *Journal of Economics & Management Strategy* 22, 2 (2013), pp. 341–364.

[20] Hsu, D. H., and Ziedonis, R. H. Patents As Quality Signals For Entrepreneurial Ventures. In: *Academy of Management Proceedings*. Vol. 2008. 1. Academy of Management. 2008, pp. 1–6.

[21] Caruana, R., Karampatziakis, N., and Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. In: *Proceedings of the 25th International Conference on Machine learning*. ACM. 2008, pp. 96–103.

[22] Mitchell, T. M. *Machine Learning*. McGraw-Hill, New York, 1997.

[23] Kotsiantis, S. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* 31, 3 (2007), pp. 249–268.

[24] Strobl, C., Malley, J., and Tutz, G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14, 4 (2009), p. 323.

[25] Kuhn, M., and Johnson, K. *Applied predictive modeling*. Springer, 2013.

[26] Chok, N. S. Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. `http://d-scholarship.pitt.edu/8056/1/Chokns_etd2010.pdf`. PhD thesis. University of Pittsburgh, 2010.

[27] Pedregosa, F., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, Oct (2011), pp. 2825–2830.

[28] Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*. Vol. 1. Springer, Berlin, 2001.

[29] Davis, J., and Goadrich, M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM. 2006, pp. 233–240.

[30] Xiang, G., et al. A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence. 2012, pp. 607–610.

[31] Ahn, H., and Kim, K.-j. Using genetic algorithms to optimize nearest neighbors for data mining. *Annals of Operations Research* 163, 1 (2008), pp. 5–18.

[32] Liang, Y. E., and Yuan, S.-T. D. Predicting investor funding behavior using crunchbase social network features. *Internet Research* 26, 1 (2016), pp. 74–100.

[33] Fried, V. H., and Hisrich, R. D. Toward a model of venture capital investment decision making. *Financial Management* (1994), pp. 28–37.

[34] Caruana, R., and Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In: *Proceedings of the 23rd International Conference on Machine Learning*. ACM. 2006, pp. 161–168.

[35] Breiman, L. Random forests. *Machine learning* 45, 1 (2001), pp. 5–32.

[36] Shin, K.-S., Lee, T. S., and Kim, H.-j. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications* 28, 1 (2005), pp. 127–135.

A Supervised Approach to Predicting the Acquisition of Startups in the Private Markets

Original Research Proposal

W.M.R. Shelton

This report is submitted as partial fulfilment
of the requirements for the Honours Programme of the
School of Computer Science and Software Engineering,
The University of Western Australia,
2017

Component: Research Proposal

Supervisors: Professor Melinda Hodkiewicz, Dr Tim French

Degree: BPhil(Hons) (24 point project)

University: The University of Western Australia

Background

High-growth technology companies (startups) are turning away from the public markets. Amazon went public in 1997, just two years after its first round of institutional financing, at a market capitalisation of $440M [1]. Contrast that with Uber, which remains private six years on and recently raised $3.5B at a $59B pre-money valuation [2]. Time to Initial Public Offering (IPO) for Venture Capital (VC)-backed startups has more than doubled over the past 20 years while VC-backed startups pursuing an IPO has plummeted [3].

One explanation for why startups are staying private for longer is the accelerating nature of global business. Startups, particularly those backed by VC firms, are expected to scale fast and require frequent rounds of fundraising coupled with centralized, quick decision making. Such flexibility is not afforded to public companies, due to strict reporting and compliance requirements [4].

Why does this waiting game matter? Principally, because it shifts value creation to the private markets. To put things in perspective, Microsofts market capitalisation grew 500-fold following its IPO [5], but for Facebook to do the same now its valuation would have to exceed the global equity market [6]. VC funding for late-stage startups is approaching all-time highs, possibly because more investors are entering the private markets to seek higher returns [3].

Merger and Acquisitions (M&A) have far surpassed IPOs as the most common liquidity event for startup founders and investors. In 2015, five times as many US-based VC-backed startups were acquired compared to those that went public through an IPO [3]. Accordingly, startup founders and investors may be interested in predicting which startups are likely to be acquired and by whom. However, M&A prediction is a challenging task.

Previous work has relied on relatively small data sets [7] because publicly-available information on private companies is scarce. In addition, previous work has focused on the financial or managerial features of potential targets [8] with little work on textual or social network features.

Xiang and colleagues [9] addressed some of these challenges by mining CrunchBase profiles and TechCrunch news articles to predict the acquisition of private startups. Their corpus was larger than previous studies: 38,617 TechCrunch news articles from June 2005 - December 2011 mentioning 5,075 companies, and a total of 59,631 CrunchBase profiles collected in January 2012. Their approach achieved a True Positive rate of between 60-79.8% and a False Positive rate of between 0-8.3%.

There are limitations to Xiang and colleagues' study: the CrunchBase corpus they studied was sparse, only a few common binary classification techniques were tested, and their approach didn't consider IPOs or bankruptcies as potential outcomes. In addition, it is unclear how robust their classifiers are through time. The study could be extended by applying the topic modelling approach to other text corpora such as patent filings, or by attempting a social network link prediction model.

Aim

We aim to produce a supervised learning model that will accurately predict the acquisition of startups in the private markets. We will build on the study by Xiang and colleagues (2012) [9], introducing new features and classification techniques. In the previous study, True Positive rate (TP), False Positive rate (FP) and Area under the ROC curve (AUC) were the main evaluation metrics used (collectively, known as "accuracy").

Hypothesis 1 (H1) Xiang and colleagues (2012) [9] results can be replicated

H2 Introducing new classification techniques improves accuracy

Xiang and colleagues' study tested three common binary classification techniques: Bayesian Networks (BN), Support Vector Machines (SVM), and Logistic Regression (LR). BN significantly outperformed SVM and LR. The authors suggested that this was because of the high correlation among their features and absence of a linear separator in the feature space. We will test a number of new classification techniques including Random Forests (RF), CART Decision Trees (CART), and Restricted Bolzmann Machines (RBM), to try to improve the accuracy of the model.

H3 Introducing additional CrunchBase features improves accuracy

Xiang and colleagues' study used a total of 22 factual features from CrunchBase profiles. No feature selection process was documented. A recent similar study on AngelList (which has a sharing agreement with CrunchBase)

used 85 features of which 11 were selected [10]. Of those 11 features, many were not included in Xiang and colleagues' model. It is plausible that broadening the feature space may result in an improved model.

H4 Introducing additional labels improves accuracy

Xiang and colleagues' study labelled startups as either "acquired" or "not acquired". The "not acquired" category thus includes startups that have bankrupted as well as highly successful startups that went public through an IPO. It is plausible that the breadth of this category would lead to misclassification. Introducing labels for "public" and "bankrupt" could improve the accuracy of the model.

H5 Using more recent CrunchBase corpora improves accuracy

Xiang and colleagues' study used a CrunchBase corpus from January 2012. They found the corpus relatively sparse at the time. Since 2012, the CrunchBase corpus has significantly grown. The CrunchBase Venture Program and the AngelList - CrunchBase data sharing agreement have contributed to the corpus, in addition to natural growth over time. It is plausible that a more recent CrunchBase corpus will provide a better basis for a more accurate model.

This study will improve our understanding of the determinants of startup acquisition in the private markets. The system devised by this study also has the potential to de-risk venture capital and encourage greater investment in private startups.

Method

1. Replicate study by Xiang et al. (2012) [9]

   We have requested access to the CrunchBase and TechCrunch datasets used in the previous study (Note: These datasets are currently available on the Carnegie Mellon University intranet). If we are unable to access these datasets we will use a CrunchBase database snapshot from December 2013.

   - Features:
     - Factual Features (CrunchBase)
       * Basic Features e.g. office location, company age
       * Financial Features e.g. investment per funding round

        * Managerial Features e.g. number of acquired companies by founders
    - Topic Features (TechCrunch articles)
  - Outcome: Acquired? (CrunchBase)
  - Processing:
    - Topic model - Latent Dirichlet Allocation (LDA)
    - Classification techniques
      * Bayesian Network (BN)
      * Support Vector Machines (SVM)
      * Logistic Regression (LR)

2. Test additional classification techniques
   - CART Decision Tree (CART) as in [10]
   - Restricted Bolzmann Machine (RBM) as in [10]
   - Random Forest (RF)
   - And other classification techniques
3. Expand the factual features set
   - Founder education (CrunchBase, Dec-2013) as in [10]
   - Founder employment (CrunchBase, Dec-2013) as in [10]
   - Founding team (CrunchBase, Dec-2013) as in [11]
   - And other factual features in the CrunchBase corpus
4. Incorporate other potential startup outcomes
   - Outcomes: Bankrupt, Acquired, Public
   - Classification techniques: One vs. all (OVA), All vs. all (AVA)
5. Test classifier robustness over different datasets
   - Original dataset from Xiang et al. (2012) [9]
   - CrunchBase readily-available snapshot (December 2013)
   - CrunchBase recent crawl (September 2016)
6. Extend topic modelling and introduce network features (stretch goal)
   - Domain-Constricted LDA model (TechCrunch articles) as in [12]

- Patent similarity (Google Patents) as in [13]
- Social network link prediction (CrunchBase) as in [14, 15]
- And other types of features as time permits

Timeline

Please see below (Table 1) for a schematic of the proposed methodology.

S:W	**Date**	**Task**
2:03	Fri 19 August	Draft proposal due
2:05	29 Aug - 02 Sep	Proposal defence to research group
2:07	Fri 09 September	Data collected
2:09	Fri 23 September	Replicated previous study
2:SB	Fri 30 September	Draft literature review due
2:12	Fri 28 October	Revised proposal due
2:12	Fri 28 October	Literature review due
2:17	Fri 02 December	Completed main experiments
1:08	Fri 28 April	Draft dissertation due
1:10	Fri 12 May	Seminar title and abstract due
1:13	Mon 29 May	Final dissertation due
1:13	Fri 02 June	Poster due
1:13	29 May - 02 June	Seminar
1:17	Mon 26 June	Corrected dissertation due

Table 1: Proposed timeline

Software and Hardware Requirements

This project will be developed primarily in Python using scikit-learn, a free open-source machine learning library [16]. MySQL may be used to prepare datasets for processing. The system will be hosted on a public compute cloud, likely Amazon Web Services. A free academic license for CrunchBase has been requested.

Bibliography

[1] Kawamoto, D. *Amazon.com IPO skyrockets*. http://www.cnet.com/au/news/amazon-com-ipo-skyrockets/. Online; accessed 07 Nov 2016. May 1997.

[2] Buhr, S. *Uber takes its most significant investment yet at $3.5 billion from Saudi Arabia*. https://techcrunch.com/2016/06/01/uber-takes-its-most-significant-investment-yet-at-3-5-billion-from-saudi-arabia/. Online; accessed 07 Nov 2016. June 2016.

[3] National Venture Capital Association *2016 National Venture Capital Association Yearbook*. http://www.nvca.org/?ddownload=2963. Online; accessed 06 Nov 2016. Mar. 2016.

[4] Wies, S., and Moorman, C. Going public: how stock market listing changes firm innovation behavior. *Journal of Marketing Research* 52, 5 (2015), pp. 694–709.

[5] McNamara, P. *If you had bought 100 shares of Microsoft 25 years ago ...* Ed. by Network World. http://www.networkworld.com/article/2228727/data-center/data-center-if-you-had-bought-100-shares-of-microsoft-25-years-ago.html. Online; accessed 06 Nov 2016. Mar. 2011.

[6] Raice, S., Das, A., and Letzing, J. *Facebook prices IPO at record value*. Ed. by Journal, T. W. S. http://www.wsj.com/articles/SB10001424052702303448404577409923406193162. Online; accessed 06 Nov 2016. May 2012.

[7] Wei, C.-P., Jiang, Y.-S., and Yang, C.-S. Patent analysis for supporting merger and acquisition (M&A) prediction: A data mining approach. In: *Workshop on E-Business*. Springer. 2008, pp. 187–200.

[8] Hongjiu, L., Huimin, C., and Yanrong, H. Financial characteristics and prediction on targets of M&A based on SOM-Hopfield neural network. In: *2007 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE. 2007, pp. 80–84.

[9] Xiang, G., et al. A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence. 2012.

[10] Beckwith, J. Predicting Success in Equity Crowdfunding. Unpublished thesis. Joseph Wharton Research Scholars. Available at `http://repository.upenn.edu/joseph_wharton_scholars/25`. 2016.

[11] Spiegel, O., et al. Going it all alone in web entrepreneurship?: a comparison of single founders vs. co-founders. In: *Proceedings of the 2013 Annual Conference on Computers and People Research*. ACM. 2013, pp. 21–32.

[12] Yuan, H., Lau, R. Y., and Xu, W. The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems* 91 (2016), pp. 67–76.

[13] Huang, J., and Zhan, S. With a Little Help of My (Former) Employer: Past Employment and Entrepreneurs' External Financing. In: *Academy of Management Proceedings*. Vol. 2015. 1. Academy of Management. 2015, p. 12050.

[14] Shi, Z., Lee, G. M., and Whinston, A. B. Towards a better measure of business proximity: topic modeling for analyzing M As. In: *Proceedings of the 15th ACM Conference on Economics and Computation*. ACM. 2014, pp. 565–565.

[15] Yuxian, E. L., and Yuan, S.-T. D. Investors Are Social Animals: Predicting Investor Behaviour using Social Network Features via Supervised Learning Approach. In: *International Workshop On Mining And Learning With Graphs*. 2013.

[16] Pedregosa, F. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* (2011).

Factors that influence startup investment

Revised Research Proposal

W.M.R. Shelton

This report is submitted as partial fulfilment
of the requirements for the Honours Programme of the
School of Computer Science and Software Engineering,
The University of Western Australia,
2017

Component: Research Proposal

Supervisors: Professor Melinda Hodkiewicz, Dr Tim French

Degree: BPhil(Hons) (24 point project)

University: The University of Western Australia

Background

Technological advances have made launching startups more accessible than ever before. Customers can be accessed easily through the Internet and launching a startup can be done from a bedroom. However, startups remain competitive and risky endeavours. Startups can be unprofitable for years so entrepreneurs look for incubators, accelerators, angel investors and venture capital firms to support them through this developmental period. Aside from funding, investors hold experience and networks that can accelerate startup growth. Investors act as scouts, able to identify the potential of new startups, and as coaches, able to help startups realise that potential [1].

Startups must convince investors to support them throughout their development, but this process can be burdensome and time-consuming. Investors find it difficult to evaluate startup potential for investment because metrics of performance often do not exist or are uncertain [2]. Popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of a desire for better methods of assessing startup potential. By 2014, over 1,200 investment organisations (including 624 venture capital firms) were members of CrunchBase's Venture Program, mining CrunchBase's startup data to help inform their investment decisions [3].

Investment comes with trade-offs for startups. The majority of venture capital-backed startups end in bankruptcy [4]. Investors are protected from these losses because the minority of their investments that are successful have outsized returns: 85% of venture capital returns come from 10% of investments [4]. Investors seek to optimise the risk-reward trade-off by pressing startups to grow rapidly, frequently raise funding rounds and make quick, centralised decisions [5]. The rapid growth demanded by investors is generally incompatible with public company structures, due to reporting and compliance requirements [6]. Accordingly, we see venture capital-backed startups delaying Initial Public Offerings (IPO). Time taken to IPO has doubled in the past 20 years [7].

Aim

Startups remaining privately-held for longer shifts value creation to the private markets. Microsoft's market capitalisation grew 500-fold following its IPO in 1986, but for Facebook to grow to the same extent since its IPO in 2012 its capitalisation would exceed the total global equity market. Investment in late-stage startups is approaching all-time highs as public market investors enter the private markets [7]. Given this situation, it is important to understand how factors that influence investment change through a startup's development. A clear gap in the academic literature exists in this area. Studies of the determinants of startup investment have common weaknesses. This study will address these weaknesses in three ways:

Larger Sample Size Previous studies are restricted in sample size. Most studies have samples of fewer than 500 startups [8, 9], or between 500 and 2,000 startups [10, 11, 12, 13, 14], and only a few have large scale samples (more than 100,000 startups) [15, 16]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [17]. Startups databases (e.g. CrunchBase) and social networks (e.g. Twitter) offer larger data sets than those previously studied. We expect data collected from these sources will lead to the discovery of additional features and higher accuracy in startup investment prediction.

Developmental Focus Prior work focuses on early-stage investment in startups, primarily equity crowdfunding [18, 8, 16, 19] and angel investing [14]. The functions and objectives of startups change through their development [20]. For example, early stages of funding are characterised by uncertainty about technical validity and market fit [21]. In this setting, patents are a strong signal to investors, but may become less so in later rounds. Generally, we expect signals that attract investment in startups will change over time.

Rich Features Prior work focuses on basic company features (e.g. the headquarters' location, the age of the company) for startup investment predictive models [18, 9]. Semantic text features (e.g. patents, media) [10, 19] and social network features (e.g. co-investment networks) [13, 16, 11] may also predict startup investment. We expect a model that includes semantic text and social network features alongside basic company features could lead to better startup investment prediction.

We will develop software that collects and processes information about startups to predict their likelihood of raising investment at different stages of their

development. This study has potential for scholarly, policy and firm-specific implications. Our scholarly contribution is a conceptual framework for startup investment, based on work by Ahlers et al. [8]. Our conceptual framework posits that startup investment is a product of two factors: startup potential and investment confidence. We will test this framework with respect to startup development using cross-sectional and longitudinal analyses. We aim to contribute to the understanding of the determinants of startup investment, with a focus on how they change over time. Ultimately, we hope that we can encourage greater investment in startups.

Method

Data Collection We will develop an automated data collection system that will provide a platform on top of which we can easily perform our analyses. Our primary data sources are CrunchBase, AngelList, Twitter and PatentsView. We will start with a focus on CrunchBase and then develop systems to match the other sources. CrunchBase data can be accessed in multiple ways. The simplest format are comma separated files (CSV) that hold data about each relation in their database (e.g. funding rounds, investors). A current CSV dump of the database can be requested at any time from CrunchBase. We have also retrieved several older CSV dumps that can be compared with current data for longitudinal studies. CSV dumps provide a subset of the attributes in the CrunchBase data set. To get the full data set requires access through their application programming interface (API). We will develop a crawler that can continually traverse the API iteratively to effectively mirror the CrunchBase data set locally for further analyses. Our master database is likely to be a Sqlite server. We are also investigating distributed solutions, including using Spark.

Machine Learning Analyses We will manipulate and combine the data collected from our data sources into a labelled data set appropriate for the application of supervised machine learning algorithms. Primary labels will be whether a startup receives funding at each funding round. We may also investigate measures of startup performance (e.g. survival time, exit). We will compare and evaluate machine learning algorithms to find which algorithms suits this task best. We have collected six historical CSV dumps from CrunchBase spanning the period from October 2013 to the present. We will match companies across these data sets to test the robustness of our model across time and to see whether the gradient of change in different features can provide greater accuracy to our model than the static features.

Timeline

Please see below (Table 1) for a schematic of the proposed methodology.

S:W	**Date**	**Task**
2:03	Fri 19 August	Draft proposal due
2:14	Wed 09 November	Revised proposal due
2:14	Wed 09 November	Literature review due
2:16	Fri 25 November	Data collected
2:21	Fri 30 December	Completed main experiments
1:08	Fri 28 April	Draft dissertation due
1:10	Fri 12 May	Seminar title and abstract due
1:13	Mon 29 May	Final dissertation due
1:13	Fri 02 June	Poster due
1:13	29 May - 02 June	Seminar
1:17	Mon 26 June	Corrected dissertation due

Table 1: Proposed timeline

Software and Hardware Requirements

This project will be developed primarily in Python using scikit-learn, a free open-source machine learning library [22]. Sqlite may be used to prepare datasets for processing. The system will be hosted on a public compute cloud, likely Amazon Web Services. Free academic licenses for CrunchBase and AngelList have been requested.

Bibliography

[1] Baum, J. A., and Silverman, B. S. Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing* 19, 3 (2004), pp. 411–436.

[2] Shane, S., and Cable, D. Network ties, reputation, and the financing of new ventures. *Management Science* 48, 3 (2002), pp. 364–381.

[3] Patil, A. *CrunchBase's Venture Program Members Are Making Startup Data Better Than Ever*. `https://info.crunchbase.com/2015/01/crunchbases-venture-program-members-are-making-startup-data-better-than-ever/`. Online; accessed 18 05 2015. Jan. 2015.

[4] Sahlman, W. *Risk and reward in venture capital*. 2010.

[5] Fried, J. M., and Ganor, M. Agency costs of venture capitalist control in startups. *New York University Law Review* 81 (2006), p. 967.

[6] Wies, S., and Moorman, C. Going public: how stock market listing changes firm innovation behavior. *Journal of Marketing Research* 52, 5 (2015), pp. 694–709.

[7] National Venture Capital Association *2016 National Venture Capital Association Yearbook*. `http://www.nvca.org/?ddownload=2963`. Online; accessed 06 Nov 2016. Mar. 2016.

[8] Ahlers, G. K., et al. Signaling in equity crowdfunding. *Entrepreneurship Theory and Practice* 39, 4 (2015), pp. 955–980.

[9] Gimmon, E., and Levie, J. Founder's human capital, external investment, and the survival of new high-technology ventures. *Research Policy* 39, 9 (2010), pp. 1214–1226.

[10] Hoenen, S., et al. The diminishing signaling value of patents between early rounds of venture capital financing. *Research Policy* 43, 6 (2014), pp. 956–989.

[11] Yu, Y., and Perotti, V. Startup Tribes: Social Network Ties that Support Success in New Firms. In: *Proceedings of 21st Americas Conference on Information Systems*. 2015.

[12] An, J., Jung, W., and Kim, H.-W. A Green Flag over Mobile Industry Start-Ups: Human Capital and Past Investors as Investment Signals. In: *PACIS 2015 Proceedings*. AIS Electronic Library, 2015, p. 67.

[13] Werth, J. C., and Boeert, P. Co-investment networks of business angels and the performance of their start-up investments. *International Journal of Entrepreneurial Venturing* 5, 3 (2013), pp. 240–256.

[14] Croce, A., Guerini, M., and Ughetto, E. Angel Financing and the Performance of High-Tech Start-Ups. *Journal of Small Business Management* (2016). issn: 1540-627X. doi: 10.1111/jsbm.12250.

[15] Shan, Z., Cao, H., and Lin, Q. Capital Crunch: Predicting Investments in Tech Companies. Unpublished thesis. Stanford University. Available at http://www.zifeishan.org/files/capital-crunch.pdf. 2014.

[16] Cheng, M., et al. Collection, exploration and analysis of crowdfunding social networks. In: *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*. ACM. 2016, pp. 25–30.

[17] Caruana, R., Karampatziakis, N., and Yessenalina, A. An empirical evaluation of supervised learning in high dimensions. In: *Proceedings of the 25th International Conference on Machine learning*. ACM. 2008, pp. 96–103.

[18] Beckwith, J. Predicting Success in Equity Crowdfunding. Unpublished thesis. Joseph Wharton Research Scholars. Available at http://repository.upenn.edu/joseph_wharton_scholars/25. 2016.

[19] Yuan, H., Lau, R. Y., and Xu, W. The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems* 91 (2016), pp. 67–76.

[20] McMullen, J. S., and Dimov, D. Time and the entrepreneurial journey: The problems and promise of studying entrepreneurship as a process. *Journal of Management Studies* 50, 8 (2013), pp. 1481–1512.

[21] Hsu, D. H., and Ziedonis, R. H. Patents As Quality Signals For Entrepreneurial Ventures. In: *Academy of Management Proceedings*. Vol. 2008. 1. Academy of Management. 2008, pp. 1–6.

[22] Pedregosa, F. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* (2011).