CHAPTER 1

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 screen many startups for investment. Current screening methods are time-consuming and subject to human selection biases. Preliminary evidence suggests that an improved system can be developed which is practical, robust and versatile.
2. Features. VC 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.

1.1 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 find that few VC firms have declared their use of data mining. We explore why the adoption of data mining in the VC industry is limited and develop criteria by which we can evaluate whether a VC investment screening system is an improvement on the status quo.

1.1.1 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 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. Reported US VC investments in 2015 totalled US$60 billion [1]. VC, 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 [2]. Compared to other forms of finance, VC financing 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 [3].

1.1.2 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 [4]. 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 [2].

Referals 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 develop investment theses for particular industries and discover companies through papers, events and databases related to those industries. These screening processes are labour-intensive and time-consuming, and even after a company is screened there is still more due-diligence needed before a final decision is made.

1.1.3 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 [5]. 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." [2].

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 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 employees, who are used to a high-degree of control in their investments. Investment screening is not considered a time-sensitive process in the VC industry (unlike structuring, for example), but screening systems should be designed so information used is always 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.” [2]. Therefore, a screening system should have minimal variance in performance when training on data sets from different times 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 lifecycles of their funds, and the portfolios that they hold [6]. 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).

1.2 Features

In this section, we provide a review of studies performed into Venture Capital (VC) investment decisions and startup performance. In Table 1.1 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 assymetry [**CITE**]. Given this context, VC investment decisions can be broken

Features	Results from Studies	
	Significant	Non-Significant
Startup Potential		
Human Capital		
Founder Capabilities	[7, 8, 9]	[10, 11]
Advisor Capabilities	[12]	[13, 8]
Executive Capabilities	[7, 8, 11]	[13]
Social Capital		
Social Influence	[7, 8, 14, 15]	-
Strategic Alliances	[12]	-
Structural Capital		
Patent Filings	[16, 17, 12]	[13, 9]
Investment Confidence		
Third Party Validation		
Investment Record	[13, 7, 18, 16, 11]	-
Investor Reputation	[8, 19, 17]	[16]
Media Coverage	[7]	[8]
Historical Performance		
Financial Performance	[7, 12]	-
Non-Financial Performance	[8, 9]	[16]
Contextual Cues		
Industry Performance	[10, 18, 9]	[7, 11]
Broader Economy	[7, 18, 16, 11, 17]	[10, 13]
Local Economy	[10, 7, 18, 9, 16]	-

Table 1.1: 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.

down into two main components: the underlying determinants of startup performance (which are difficult to observe), and signals that correlate with startup performance (which are easier to observe). We will discuss these two components at a high-level in the following sections (and in detail in Appendix ??).

1.2.1 Startup Potential

Determinants of startup performance from the literature can be broadly categorised into three areas: human capital, social capital and structural capital [12, 13].

- Human Capital. Human capital is critical to early-stage startups that have limited resources and are changing constantly. The education background of founders [7, 9], and the past entrepreneurial experience of advisors [12] and founders [9] have been linked to startup performance, though some studies dispute this [10, 11].

- Social Capital. Entrepeneurship revolves around opportunity discovery and realisation through the medium of social networks [20]. Presence and engagement on Facebook and Twitter are predictive of startup performance [14, 7] and strategic alliances have also been found to predict VC investment [12].

- 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 biotechnolgoy startups [12, 16] but there is less supportive evidence for non-biotechnology startups [9, 13].

1.2.2 Investment Confidence

A key challenge of the VC investment process is informational asymmetry [**CITE**]. 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. By their nature, startups and founders are optimsitic about the effectiveness of new technologies and business models but third party validation from credible sources like other investors [13, 7, 18, 16, 11]., the media [7], and the government, has been shown to predict investors' decision-making process.

- 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. Most studies use simplistic performance metrics like survival time [21, 22, 23] but that is not a particularly helpful measure for VC investment screening.

- Contextual Cues. Startups do not exist in isolation but are rather a product of their context. Investors consider the performance of a startup's competitors [10, 18, 9], their local economy [7, 18, 9] and the broader economy [18, 16] when evaluating potential investment candidates.

1.2.3 Feature Evaluation

In previous sections, we collected evidence of features that influence startup performance and VC investment decisions. We found that previous studies have typically focused on factors in isolation and rarely evaluate 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).

1.3 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 [13, 9] or between 500 and 2,000 startups [16, 15, 8, 19, 18]. Few studies have used larger samples (more than 100,000 startups), usually derived from CrunchBase or AngelList [10, 14]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [24]. Startup databases (e.g. CrunchBase) and social networks (e.g. Twitter) offer data sets 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 1.2, 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 this task, review potentially relevant data sources, and ultimately determine which data sources are most likely to suit the characteristics of this task.

1.3.1 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.

1.3.2 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 review, 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

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 1.2: 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.

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.

1.4 Classification Algorithms

Predicting startup performance is a difficult problem for humans and most Venture Capital (VC) investments fail. 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 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 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." [25].

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 1.3, 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 this task.

1.4.1 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 uses measures like whether a company has been

Criteria	Machine Learning Algorithms													
	NB		LR		KNN		DT		RF		SVM		ANN	
Data Set Properties	**2**		4		6		**2**		**1**		4		6	
Missing Values	✓✓	[26]	✓	-	✗	[26]	✓✓	[26]	✓✓	[27]	✓	[26]	✗	[26]
Irrelevant Features	✗	[26]	✗	[28]	✓	[26]	✓✓	[26]	✓✓	[27]	✗	[26]	✗	[26]
Imbalanced Classes	✓✓	-	✓✓	-	✗	-	✗	[26]	✓	[27]	✓✓	[26]	✓	[26]
Algorithm Properties	**2**		**1**		4		4		**2**		6		6	
Predictive Power	✗	[24]	✓	[24]	✓	[24]	✗	[26]	✓✓	[24]	✓✓	[24]	✗✓	[24]
Interpretability	✓✓	[26]	✓✓	[28]	✗	[26]	✓✓	[26]	✓	[28]	✗	[26]	✗	[26]
Processing Speed	✓✓	[26]	✓✓	[24]	✓✓	[26]	✓	[26]	✓	[24]	✗	[26]	✗	[26]
Overall	**2**		**2**		6		4		**1**		5		7	

Table 1.3: 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 data set 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.

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 characteristics of supervised learning tasks relevant to VC investment screening.

1.4.1.1 Data Set Properties

While data sets can be pre-processed to assist with their standardisation, some types of data sets 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 due to errors in data collection or processing. Missing data has different effects depending on its distribution through the data set. Public data sets, 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 data sets 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 data sets are likely to be highly imbalanced because very few startups are successful. Therefore, robustness to imbalanced classes is a desirable property of our algorithm.

1.4.1.2 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. In this case, processing speed is not critical because generally VC investment decisions are made over weeks and months, though there is some need for the data set to be updated with new information as it becomes available.

1.4.2 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.

1.4.3 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 [24]. 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 due to 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 data set.

1.5 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 [13, 9, 29, 16, 15, 8, 19, 18], a focus on early-stage investment [7, 13, 14, 30, 18, 2], and a basic feature set [13, 8, 14, 18, 19, 9]. 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.

APPENDIX A

Data Sources

A.1 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. 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 and AngelList provide free Application Program Interfaces (API) for academic use. Crawlers can be developed to traverse these APIs and collect data systematically. The advantages of crawlers are that they can selectively collect data from nodes with specific attributes, collect random samples, or traverse the data source indefinitely, updating entries as new data becomes available. CrunchBase also provides pre-formatted database snapshots which allows easier access to the data set. 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. Both CrunchBase and AngelList also have error-checking provisions (including machine reviews and social authentication) to prevent and remediate inaccurate entries but there is still a greater chance for error. Comparing CrunchBase and AngelList, CrunchBase tends to have more comprehensive records of funding rounds [14] and media coverage but AngelList also has a social network element where users can ‘follow each other - in a similar way to Twitter.

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. 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 [31], making it difficult to use for social network analyses. 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 [32].

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. 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 [33]. 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). 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) [34].

APPENDIX B

Classification Algorithms

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 [26]. However, Naive Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features [35]. 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 [7]. 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 [36], 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 [24]. This underperformance is observed when Logistic Regression is applied to startup investment prediction tasks [7, 37]. 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 [9].

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 [26]. 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 [38]. 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 [26]. 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 [39]. 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 [40]. 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 [41]. 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 [39]. 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 [37]. 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 [41].

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 [39], 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 [42]. 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 [43]. 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." [7]. However, this study will address these issues so Artificial Neural Networks may be more competitive.

Bibliography

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

[2] Stone, T. R. Computational analytics for venture finance. Unpublished Ph.D. dissertation. UCL (University College London), 2014.

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

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

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

[6] Gompers, P. A. Optimal investment, monitoring, and the staging of venture capital. *The journal of finance* 50, 5 (1995), 1461–1489.

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

[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.

[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), 1214–1226.

[10] 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.

[11] 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), 341–364.

[12] 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), 411–436.

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

[14] 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, 25–30.

[15] 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.

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

[17] 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, 1–6.

[18] Croce, A., Guerini, M., and Ughetto, E. Angel Financing and the Performance of High-Tech Start-Ups. *Journal of Small Business Management* (2016).

[19] 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), 240–256.

[20] Shane, S., and Venkataraman, S. The promise of entrepreneurship as a field of research. *Academy of Management Review* 25, 1 (2000), 217–226.

[21] Raz, O., and Gloor, P. A. Size really matters-new insights for start-ups' survival. *Management Science* 53, 2 (2007), 169–177.

[22] Song, Y., and Vinig, T. Entrepreneur online social networks–structure, diversity and impact on start-up survival. *International Journal of Organisational Design and Engineering* 2, 2 (2012), 189–203.

[23] Gloor, P. A., et al. Choosing the right friends–predicting success of startup entrepreneurs and innovators through their online social network structure. *International Journal of Organisational Design and Engineering* 3, 1 (2013), 67–85.

[24] 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, 96–103.

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

[26] Kotsiantis, S. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* 31, 3 (2007).

[27] 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), 323.

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

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

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

[31] Trachtenberg, A. *Changes to our Developer Program*. Ed. by Linkedin.com. `https://developer.linkedin.com/blog/posts/2015/developer-program-changes`. Online; accessed 18 05 2015. Feb. 2015.

[32] Puschmann, C., and Burgess, J. The politics of Twitter data (2013).

[33] Schultz, L. A. Preliminary Patent Searches: New and Improved Tools for Mining the Sea of Information. *Colo. Law.* 45 (2016), 55.

[34] Artemchik, T. PrivCo. *Journal of Business & Finance Librarianship* 20, 3 (2015), 224–229.

[35] Niculescu-Mizil, A., and Caruana, R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, 625–632.

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

[37] Bhat, H., and Zaelit, D. Predicting private company exits using qualitative data. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Huang, J., Cao, L., and Srivastava, J. Vol. 6634. Lecture Notes in Computer Science. Springer, Berlin, 2011, 399–410.

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

[39] 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, 161–168.

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

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

[42] 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), 127–135.

[43] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Cognitive Modeling* 5, 3 (1988), 1.