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 privately-collected datasets, a focus on early-stage investment, and limited practical application. A multi-stage supervised classification system was developed that identifies startup companies likely to receive additional funding, be acquired or have an IPO in the future. The system produces an optimised classification pipeline which is applied to data collected from large, public online databases (CrunchBase and PatentsView). The system was evaluated against three criteria important to the VC industry: predictive power, robustness, and efficiency. The system satisfies each of these criteria. It makes predictions with an average F1 score (for the positive class) of 0.36-0.55 over forecast windows of 2-4 years, it is robust, with only minimal variance in performance when trained on different historical datasets, and it is efficient in that it is semi-autonomous and could be made fully autonomous with minimal development. The prior experience of a startup's advisors, executives and founders was found to be the greatest predictor of investment performance. This project lays the groundwork for an industry-ready VC investment screening tool.

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

Acknowledgements

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 startup companies. VC firms are behind many successful high-tech companies, such as Google, Apple, Microsoft and Alibaba. VC firms fund startup companies with cash in exchange for an equity stake but 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, able to signal the potential of new startups, and as coaches, able to help startups realise that potential [5]. In this way, the experience, skills and networks of the VC firms who invest in a startup directly influence the startups trajectory. The widespread adoption of the Internet and inexpensive, ubiquitous computing has decreased the cost of starting a business and transformed the venture funding landscape – companies require less funding to launch but require more to scale in highly competitive markets [22]. There is now an impetus for VC firms to change the way they do business.

VC firms face the challenge of choosing a few outstanding investments from a sea of hundreds of thousands of potential opportunities. VC firms seek 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 commonly referred to as an 'exit) is typically either an Initial Public Offering (IPO) or an acquisition by a larger competitor. Most VC firms expect their investments to reach a liquidity event within 3–8 years, in accordance with their fund timeframe [**CITE**]. When compared to traditional investors, this would be considered a long-term investment strategy. However, when compared to the trajectory of most companies, this period is particularly short – not many companies, even successful companies, are capable of maturing at this rate. This makes the VC investment screening process particularly difficult. In addition, there are many potential investment candidates, most of which ultimately fail, and traditional metrics of performance (e.g. cashflow, earnings) often do not exist or are uncertain [42]. Traditionally, investment opportunities are either referred or identified through technology scans (e.g. Google searches, patent searches). These manual search processes are time-consuming for VC firms.

The VC industry requires better systems and processes to efficiently manage labour-intensive tasks like investment origination and screening. Existing approaches in the literature to predict startup performance have three common limitations: small sample size [1, 20, 16, 24, 52, 3, 50, 14], a focus on early-stage investment [6, 1, 12, 53, 14, 46], and sparse use of features [1, 3, 12, 14, 50, 20]. Although individual studies address some of these limitations, none attempt to synthesise their findings into a standalone study and software design. In addition, there is little evidence that previous research has been translated into systems that are able to assist investors directly. The popularity of online databases like AngelList and CrunchBase, which offer information on startups, investments and investors, is evidence of the VC industrys desire for better, more quantitative methods of assessing startup potential. By 2014, over 1,200 investment organisations (including 624 VC firms) were members of CrunchBase's Venture Program, mining CrunchBase's startup data to help inform their investment decisions [33]. There is preliminary evidence that mining these data sources may address previous limitations and make investment origination and screening more efficient and effective [46, 7].

We believe it is now possible to address previous limitations in this domain and produce a VC investment screening system that is efficient, robust and powerful. Our system is based around identifying startup companies that are likely to receive additional funding or have a liquidity event (exit) in a given forecast window. This system can generate statistics and make recommendations that may assist VC firms to efficiently and effectively screen investment candidates. To be useful in this context, the implementation of the system must meet the following criteria:

1. Efficiency. Our system must be more efficient than traditional, manual investment screening. A technique to achieve this is autonomously collecting data from readily-available, online data sources. In this project, we focus primarily on the CrunchBase online database. We test whether this source provides enough observations to provide meaningful statistics.We also test whether we can produce a comprehensive and generalisable feature set, one which would allow investors to complement or replace their data sources over time.

2. Robustness. Our system must be robust enough to be reliable over time and agnostic to specific data sources. The system must provide a generalised, robust solution for investors that does not require significant technical knowledge to use, is not specific to a time-period, and is not reliant on

a single data source. The parameters and features that the system selects should be invariant to time so investors can have reasonable confidence in its predictions.

3. Predictive Power. Our system must be consistently accurate at identifying a variety of high-potential investment candidates. The system provides a broad first screening process for investors so it is important that it is highly sensitive (i.e. it excludes very few positive results). The system should be robust to different forecast windows (i.e. exit in three years from now) as VC firms make investment decisions with different periods so they can strategically manage the investment horizons of their funds. Similarly, the system should accurately assess companies at any developmental stage.

The following work is presented in three chapters:

1. Literature Review. We review the theoretical background of startup performance and VC investment and evaluate previous attempts at using data mining in this domain. We determine that the VC industry requires better systems to efficiently manage labour-intensive tasks like investment screening. Existing approaches in the literature to tackle similar business problems have three common limitations: small sample size, a focus on early-stage investment, and incomplete use of features. Preliminary evidence suggests that online data sources and machine learning techniques may allow us to address previous limitations and produce an investment screening system that is efficient, robust and powerful.

2. Design. We outline the design of our system architecture. Our system uses data from the CrunchBase online database with some supplementation from PatentsView (US Patents Office). We use two datasets collected in September 2016 and April 2017 for training and testing respectively. We develop a classification pipeline using the popular Python-based machine learning library Scikit-learn [34]. Our pipeline includes imputation, feature transformation and scaling, extraction and classification. We optimised these steps using cross-validated randomized hyperparameter search.

3. Evaluation. We evaluate our proposed system against three criteria: efficiency, robustness and predictive power. Firstly, we evaluate efficiency by exploring the learning curves of our classification techniques and whether there is sufficient data to produce reliable statistics. Secondly, we evaluate robustness by evaluating our models against multiple reverse-engineered historical datasets and measuring their variance. Thirdly, we evaluate predictive power by testing different forecast windows outcomes. Finally, we

discuss our findings more broadly and their implications for investors and future research into startup investment and performance.

CHAPTER 2

Literature Review

In this chapter, we first review the startup investment literature to develop criteria to evaluate our Venture Capital (VC) investment screening system. We then turn our focus to determining the best techniques to use to create this system, which we break down into three intercorrelated areas: feature selection, data sources and classification algorithms.

1. Criteria Selection. VC firms review many potential investment candidates to shortlist for investment. Traditional screening methods involve referral, networking and Internet search. These screening methods are highly time-consuming and subject to human selection biases. Based on our review, we believe that a superior system can be produced and should be assessed on the basis of its efficiency, robustness, and predictive power.

2. Feature Selection. VC is a key driver of startup development but our understanding of factors that influence VC firms' investment decisions and the subsequent performance of those investments is incomplete. Based on our review of the literature, we propose a hierarchical framework that includes a variety of features that have been previously indicated to be relevant to this investment screening problem. At a high-level, our framework incorporates determinants of startup potential and signals that influence investment confidence.

3. Data Sources. Startup performance is a multi-faceted problem and different data sources provide insights into different actors, relationships and attributes. Our review focuses on novel online data sources which have the potential to transform entrepreneurship and VC research. Preliminary evidence suggests that the online startup databases CrunchBase and AngelList are promising and likely to provide a comprehensive feature set that can form the basis of our system. Other sources like PatentsView, Twitter, LinkedIn, and PrivCo are considered.

4. Classification Algorithms. Predicting startup performance is a difficult problem for humans. After all, a high percentage of even VC-backed startups still fail. However, machine learning techniques have been recently used in other areas of finance (e.g. in the public markets) with some success. We cross-reference the characteristics of our intended dataset with the characteristics of common supervised classification algorithms. Our analyses suggest that we should expect Random Forests, Support Vector Machines and Artificial Neural Networks to be most suitable for our system.

2.1 Criteria Selection

Venture Capital (VC) financing has lagged behind other forms of high finance (e.g. bond trading, loan applications, insurance) in adopting computational analytics to aid decision-making. Banks are now able to evaluate personal 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 venture finance is still foreseeable.

In this section, we provide an introduction into VC firm strategy and review the existing state of the VC investment process. We find that analytical tools are nascent and use of analytics in industry is limited. To date only a small handful of VC firms have publicly declared their use of computational analytical methods in their decision making and investment selection process. We explore why the use of data mining in the VC industry is limited and we develop criteria by which we can judge a VC investment screening system to be successful.

2.1.1 Venture Capital Industry

Early-stage investment is a key driving force of technological innovation and is vitally important to the wider economy, especially in high-growth 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 [31]. Reported US VC investments in 2015 totalled US$60 billion [31].

Figure 2.1 illustrates the typical structure of a VC Fund. A typical fund is managed by a VC Firm (legally referred to as a General Partner) consisting of several investment partners. The fund itself (the Limited Partnership) is essentially an investment fund raised from various institutional investors such as pension funds, university endowments and family offices (legally referred to as Limited Partners). Beyond fundraising the main responsibilities of investment partners (also referred to as General Partners) are sourcing investment opportunities, making investment decisions and taking board membership to assist the management of investee private companies (also referred to as Portfolio Companies).

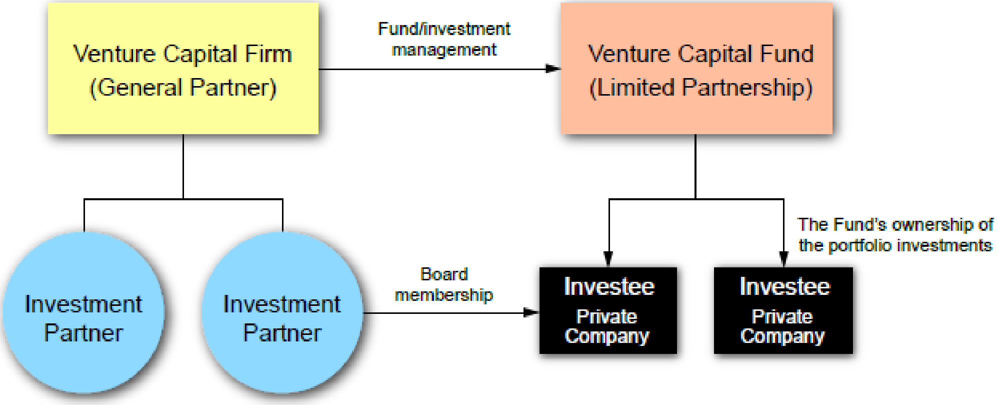


Figure 2.1: Venture Capital fund structure.

2.1.2 Venture Capital Firm Strategy

Typically, VC firms are reliant on a small number of high-risk investments to produce outsized returns through successful exit events. A common rule-of-thumb is that given a portfolio of ten startup companies: three will fail entirely, three will remain active but will not be very profitable, three will be active and profitable, and one highly successful startup will provide the investor with a multiple return on all of the investments [46]. In comparison to other traditional investment classes, VC financing is heavily biased towards control at the expense of risk mitigation. Although VC firms tend not to take majority stakes in startups, they exert their influence through significant minority stakes, board membership, their relative seniority to the companys founders, and through leveraging their business networks [17].

Despite VC firms, often significant, influence on the trajectory of their investments, they are still highly selective of the companies that they invest in. Although rarely reported, a small number of studies show VC investment rates vary between 1.5-3.5% of proposals considered [46]. Accordingly, traditional venture finance is a very labour intensive and time consuming process involving extensive due diligence on behalf of the investor [18]. The VC investment process involves several main stages: deal origination, screening, evaluation, structuring (e.g., valuation, term sheets), and post investment activities (e.g., recruiting, financing).

2.1.3 Current Venture Capital Systems

Early-stage investment 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 a high time-cost of undertaking due diligence. This makes for a complicated origination and screening process. While referral from trusted sources (e.g., entrepreneurs, accountants, lawyers, other investors) is often used to screen opportunities, as the cost of starting businesses dramatically decreases investors are faced with an increasingly large number potential businesses and investment opportunities to assess and evaluate. Such a proliferation has led to an information overload problem in venture finance.

Despite evidence that VC firms could benefit from increased use of data mining, it appears few are interested in advanced data analytics. Stone [46] interviewed Fred Wilson of Union Square Ventures who said: “We have not been able to quantify [startup potential]. We havent 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.” Likewise, when asked, Chris Dixon of Andreessen Horowitz said: “Ive 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 overfitting (finding patterns in the past that dont carry forward in the future”.

Similarly, while recently new software tools have been developed to assist VC firm, there is limited evidence of their adoption.

2.1.4 Proposed Criteria

Based on our review of the VC industry and current VC origination and screening processes, we have developed criteria on which we can evaluate our proposed system.

1. Efficiency. Our system must be more efficient than traditional, manual investment screening by referral and technology scan (e.g. Google search, media, databases). This means that it needs to be able to provide enough

information – both observations and features – to be able to meet similar levels of accuracy.

2. Robustness. Our system must be robust enough to be reliable over time. The system must provide a generalised, robust solution for investors that does not require significant technical knowledge to use, and is not overfitted to a specific time-period or data source.

3. Predictive Power. Our system must be consistently accurate at identifying a variety of high-potential investment candidates. The system should be robust to different forecast windows (i.e. exit in three years from now) as VC firms make investment decisions with different periods so they can strategically manage the investment horizons of their funds.

2.2 Feature Selection

Our understanding of the factors that influence Venture Capital (VC) investment decisions and the subsequent performance of those investments is incomplete. We believe a diverse range of features is critical to developing accurate models of startup performance and investment decisions.

Prior work focuses on basic company features (e.g. the headquarters' location, the age of the company) for startup investment predictive models [6, 20]. Semantic text features (e.g. patents, media) [24, 53] and social network features (e.g. co-investment networks) [50, 12, 52] 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 propose a conceptual framework that builds upon previous work to ensure that we include a comprehensive and relevant set of features in our investment screening system. Ahlers et al. [1] developed a conceptual framework for funding success on equity crowdfunding platforms. Their framework has two factors: venture quality and level of uncertainty. The first factor is based on work by Baum and Silverman [5] that suggests key determinants of startup potential are human capital, alliance (social) capital, and intellectual (structural) capital. The second factor is based on investors' confidence in their estimation of startup potential.

We seek to generalise Ahlers' framework [1] beyond equity crowdfunding. While the first factor of Ahlers' framework (venture quality) applies to startups of all stages, Ahlers operationalise their second factor with respect to whether startups offer an equity share in their crowdfunding, and whether they provide financial projections. These features are specific to equity crowdfunding. We propose an extension of Ahlers' framework that generalises and develops this second factor. We describe investment confidence as a product of third party validation, historical performance and contextual cues. Our proposed framework is depicted in Figure 2.2.

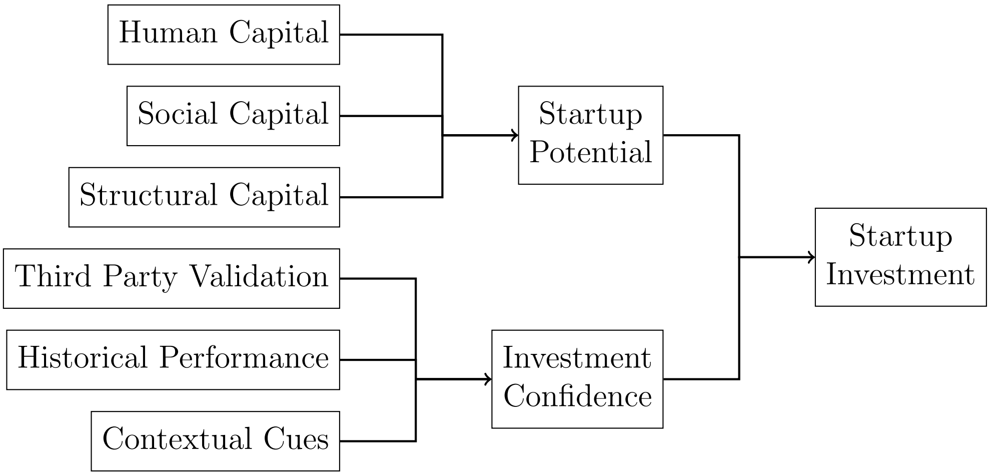


Figure 2.2: Proposed conceptual framework for startup investment. We adapt the framework proposed by Ahlers et al. [1], originally based on work by Baum and Silverman [5]. For an extended version of this framework, please refer to Figure A.1.

Next, we must operationalise this conceptual framework into features that we can incorporate into our machine learning model. Table 2.1 shows a review of features tested in previous studies of startup investment. In Appendix A, we describe each of these features and outline theoretical and empirical evidence that justify their inclusion in our conceptual framework.

Feature selection is critical to the success of our proposed conceptual framework. In this section, we have built on the framework proposed by Ahlers et al. [1] in several ways. First, our framework generalises the "Investment Confidence" factor for startups seeking any type of investment (not just equity crowdfunding). Second, our framework has greater depth. Where Ahlers uses one or two features for each factor in their model (e.g. "% Nonexecutive board" represents "Social (alliance) capital"), we perform a review of many features employed in this area and perform a higher degree of classification. For example, in our proposed framework "Social (alliance) capital" is composed of "Social influence" and "Strategic alliances", each of which will also be composed of several features (e.g. "Twitter followers", "Average Tweets per day").

Features	Results from Studies	
	Significant	Non-Significant
Startup Potential		
Human Capital		
Founders' Capabilities	[6, 3, 20]	[41, 13]
NED Capabilities	[5]	[1, 3]
Staff Capabilities	[6, 3, 13]	[1]
Social Capital		
Strategic Alliances	[5]	-
Social Influence	[6, 3, 12, 52]	-
Structural Capital		
Patent Filings	[24, 25, 5]	[1, 20]
Investment Confidence		
Third Party Validation		
Investment Record	[1, 6, 14, 24, 13]	-
Investor Reputation	[3, 50, 25]	[24]
Media Coverage	[6]	[3]
Awards and Grants	[1]	-
Historical Performance		
Financial Performance	[6, 5]	-
Non-Financial Performance	[3, 20]	[24]
Contextual Cues		
Competitor Performance	[41, 14, 20]	[6, 13]
Broader Economy	[6, 14, 24, 13, 25]	[41, 1]
Local Economy	[41, 6, 14, 20, 24]	-

Table 2.1: Features relevant to startup investment. We review thirteen empirical studies that investigate drivers of startup investment. For each study, we note whether included features have a significant effect on the startup investment model. We classify identified features according to our proposed conceptual framework.

2.3 Data Sources

Predicting startup investment and performance is a complex and difficult task. There are many features that can influence startup investment decisions. Capturing the diversity of these features is critical to developing accurate models. Accordingly, this task will likely involve data collection from multiple data sources. Appropriate selection of these data sources is important because different data sources provide insights into different actors, relationships and attributes.

Previous studies in this field have been limited by data sources restricted in sample size. Most studies have samples of fewer than 500 startups [1, 20], or between 500 and 2,000 startups [24, 52, 3, 50, 14], and only a few have large scale samples (more than 100,000 startups) [41, 12]. Sample size is more critical to model development than the sophistication of machine learning algorithms or feature selection [10]. 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.

In Table 2.2, we outline the characteristics of relevant data sources and how they could contribute to our chosen features. In this section, 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.

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). In the following section we review the characteristics of each of these data sources commonly used in entrepreneurship research.

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

Table 2.2: Data sources relevant to startup investment. We review six data sources commonly used in entrepreneurship research for their suitability for our startup investment task. We evaluate data sources for their ability to provide relevant features for our analyses and for their ease of use in data collection. We exclude offline sources from our analyses. Ratings are: ✗ = poor, ✓ = satisfactory, ✓✓ = good.

2.3.1.1 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 [12] and media coverage but AngelList also has a social network element where users can 'follow each other - in a similar way to Twitter.

2.3.1.2 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 [49], 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 [36].

2.3.1.3 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 [40]. 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) [4].

2.3.2 Source Evaluation

Entrepreneurship and Venture Capital (VC) research is primed to take advantage of the availability 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. There are small differences between the features recorded by each. CrunchBase has slightly more coverage and tracks media better 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 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. Computational analytics have been heavily deployed in high finance and we believe there is scope for applying related techniques to improve upon investment decision making in the domain of venture finance. 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.” [30].

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 predicting startup investment. 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 the startup investment prediction task, review common machine learning algorithms, and determine which algorithms are most likely to suit the characteristics of this task.

Criteria	Machine Learning Algorithms													
	NB		LR		KNN		DT		RF		SVM		ANN	
Data Set Properties	**2**		4		6		**2**		**1**		5		7	
Missing Values	✓✓	[27]	✓	-	✗	[27]	✓✓	[27]	✓✓	[47]	✓	[27]	✗	[27]
Mixed Feature Types	✓✓	[27]	✓✓	-	✓✓	[27]	✓✓	[27]	✓	[47]	✓	[27]	✓	[27]
Irrelevant Features	✗	[27]	✗	[28]	✓	[27]	✓✓	[27]	✓✓	[47]	✗	[27]	✗	[27]
Imbalanced Classes	✓✓	-	✓✓	-	✗	-	✗	[27]	✓	[47]	✓✓	[27]	✓	[27]
Algorithm Properties	**2**		**1**		5		5		**2**		5		**2**	
Predictive Power	✗	[10]	✓	[10]	✓	[10]	✗	[27]	✓✓	[10]	✓✓	[10]	✗✓	[10]
Interpretability	✓✓	[27]	✓✓	[28]	✗	[27]	✓✓	[27]	✓	[28]	✗	[27]	✗	[27]
Incremental Learning	✓✓	[27]	✓✓	-	✓✓	[27]	✓	[27]	✓	-	✓	[27]	✓✓	[27]
Overall	**2**		**2**		6		4		**1**		5		6	

Table 2.3: Evaluation of machine learning algorithms for startup investment prediction. We review seven common supervised machine learning algorithms for their suitability for our startup investment task. We evaluate algorithms for their robustness to the structure of the data set and their appropriateness for the constraints of our implementation. We rank the algorithms according to the sum of these measures (in each section and overall) and bold 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.

2.4.1 Task Characteristics

Machine learning tasks are diverse. Our investigation into startup investment is a task that suits supervised machine learning algorithms. We will manipulate the data we collect into a single labelled data set. Startups will be labelled based on whether they are acquired or have had an IPO at a later time. 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 our startup investment prediction task.

2.4.1.1 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. 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. 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 our data set because our proposed framework includes features that have not been thoroughly tested in the literature. Therefore, robustness to irrelevant features is a desirable property of our algorithm. 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. Our data set is not dramatically imbalanced overall, but when looking at funding status for different funding rounds it is significantly imbalanced. 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 is the ability of a machine learning algorithm to correctly classify new observations. Predictive power can be evaluated in many ways. As our data set is likely to have an imbalanced class distribution, we will evaluate predictive power based on balanced metrics like Area under the Receiver-Operator Curve and the F1 Score. 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 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 model interpretability. Some models are a "black box" in the sense that data comes in and out but the model cannot be interpreted. For this study, it is a key objective that we improve our understanding of the determinants of startup investment. Therefore, interpretability is a desirable property of our algorithm. 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 Venture Capital (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.

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 the following section, we describe each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in applications relevant to startup investment.

2.4.2.1 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 [27]. However, Naive Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features [32]. 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 [6]. 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.

2.4.2.2 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 [19], 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 [10]. This underperformance is observed when Logistic Regression is applied to startup investment prediction tasks [6, 7]. 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 [20].

2.4.2.3 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 [27]. 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 [2]. 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.

2.4.2.4 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 [27]. 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 [11]. 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 [29]. Decision Trees had the highest accuracy and are desirable because their reasoning is easily communicated to startups.

2.4.2.5 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 [8]. 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 [11]. 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 [7]. 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 [8].

2.4.2.6 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 [11], 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 [44]. Support Vector Machines outperformed Artificial Neural Networks, possibly because of the small data set.

2.4.2.7 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 [38]. 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 algorithmsartificial neural networks, Restricted Bolzmann machines, for instancecould be tried on the data set, but marginal improvements would likely result." [6]. However, this study will address these issues so Artificial Neural Networks may be more competitive.

2.4.3 Algorithm Evaluation

We evaluated supervised learning algorithms for their suitability in startup investment prediction. While our evaluation gives us directionality of fit, we hesitate to discard algorithms based on our literature review. Algorithm selection is complex and preliminary testing will provide clarity as to which algorithms should be used. In addition, larger training sets and good feature design tend to outweigh algorithm selection [10]. With those concessions in mind, our findings suggest we expect Random Forests, Support Vector Machines and Artificial Neural Networks to produce the 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.

2.5 Research Gap

The Venture Capital (VC) industry requires better systems and processes to efficiently manage labour-intensive tasks like investment screening. Existing approaches in the literature to predict startup performance have three common limitations: small sample size, a focus on very early stage investment, and incomplete use of features. 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 how to address these limitations and produce a system that will assist VC firms in originating and screening investment candidates.

Firstly, we reviewed the business problem and developed three criteria that will help us evaluate our system: efficiency, robustness and predictive power. Secondly, we developed a conceptual framework of predicting startup performance that incorporates determinants of startup potential and signals that influence investment confidence. This framework informs our feature selection. We then assessed potential data sources and found preliminary evidence that suggests that the startup databases CrunchBase and AngelList are promising and likely to provide a comprehensive feature set that can form the basis of our system. Finally, we reviewed supervised machine learning techniques 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.

Based on this literature review, we believe it is now possible to address previous limitations in this domain and produce an investment screening system that is efficient, robust and powerful. In the next chapter, we will outline the process by which we attempt to develop that system.

CHAPTER 3

Design

In this chapter, we explain the methodology used to fill the research gap identified in Chapter 2, thereby producing a system that identifies high-potential startup companies that are likely to receive additional funding or a exit in a given forecast window. Figure 3.1 depicts the overall system architecture, which is described in the following chapter. We evaluate the performance of this system in the next chapter, Chapter 4.

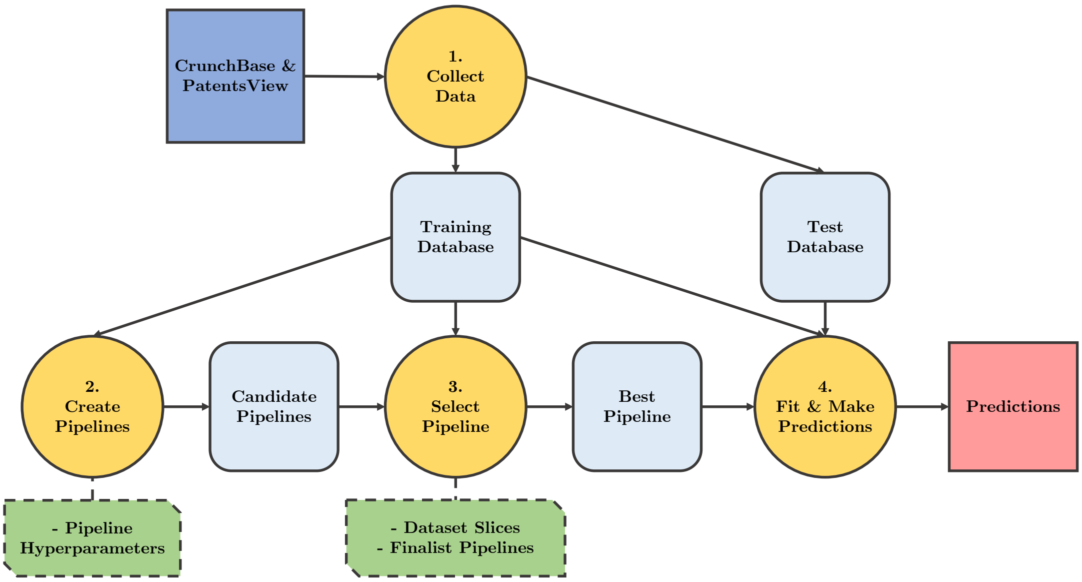


Figure 3.1: An overview of the system architecture proposed by this project, structured in four stages: data collection, pipeline creation, pipeline selection and prediction.

1. Dataset Preparation. Our primary data source was the CrunchBase online database, which we supplemented with patent filing data from PatentsView. We collected two database dumps from CrunchBase in September 2016 and

April 2017, for training and testing respectively. The database dumps were in the format of CSV files which we imported into a relational database (SQLite) and performed aggregation queries to create features suitable for classification. We then performed screening based on each company's developmental stage and age to ensure only relevant companies were included in the dataset. Finally, we explored the dataset and identified issues of sparsity, long-tailed distributions, imbalanced feature ranges, and non-orthogonality.

2. Pipeline Creation. We developed a processing pipeline framework that seeks to address the dataset issues identified during data collection. Our pipeline is based on the popular Python machine learning library Scikit-learn [34]. Pre-processing steps include imputation, transformation, scaling and feature extraction. Each pre-processing step has hyperparameters that can be tuned (e.g. imputation strategy, number of components to extract) that affect the pipeline's classification performance. We also tested a number of common classification algorithms and their hyperparameters, selected from our literature review in Chapter 2. We performed a search across the pipeline's hyperparameters to generate candidate pipelines. The hyperparameters that we found to have the most significant effect on the final performance of the pipelines were related to the classification algorithms.

3. Pipeline Selection. Our system evaluates and ranks the candidate pipelines and tests the best pipelines (finalist pipelines) over pseudo-historical datasets. This process ensures we select pipelines that are robust with respect to time. We prepared the dataset slices using a technique that filters records by their timestamps, effectively recreating historical datasets. We use Area under Precision-Recall (PR) Curve as our evaluation metric. We aggregate the results for each finalist pipeline across these dataset slices and rank the finalist pipelines on their overall performance, so we can select the best pipeline for further evaluation. We don't observe significant variance in the pipelines on aggregate against the dataset slices, but there is variance within the individual pipelines. Our results suggest that it is optimal to evaluate the top 3-5 candidate pipelines in this manner.

4. Model Fit and Prediction. Finally, our system applies the best pipeline to a training dataset to produce a fitted model. The model is then applied to a feature vector from a held-out test database, which generates a set of predictions which could, in practice, then be used by Venture Capital (VC) firms. We evaluate the accuracy of the models produced by our system with respect to a number of variables in the next chapter, Chapter 4.

3.1 Dataset Preparation

In the previous chapter, we reviewed the literature concerning data sources for entrepreneurship and Venture Capital (VC) investment. We concluded the most promising primary data sources for this project are CrunchBase and AngelList, for their size, comprehensiveness and ease of access. We suggested PatentsView (the online database of the US Patent Office) could be a useful secondary data source for structural capital features. In the following sections, we first discuss how we collected data from CrunchBase and PatentsView, converted the relational databases into a format suitable for machine learning, performed preliminary screening to ensure we only included relevant companies. This process is depicted in Figure 3.2. Following our description of this process we describe the results of exploratory analysis on our dataset and identify dataset issues which will be addressed in later steps.

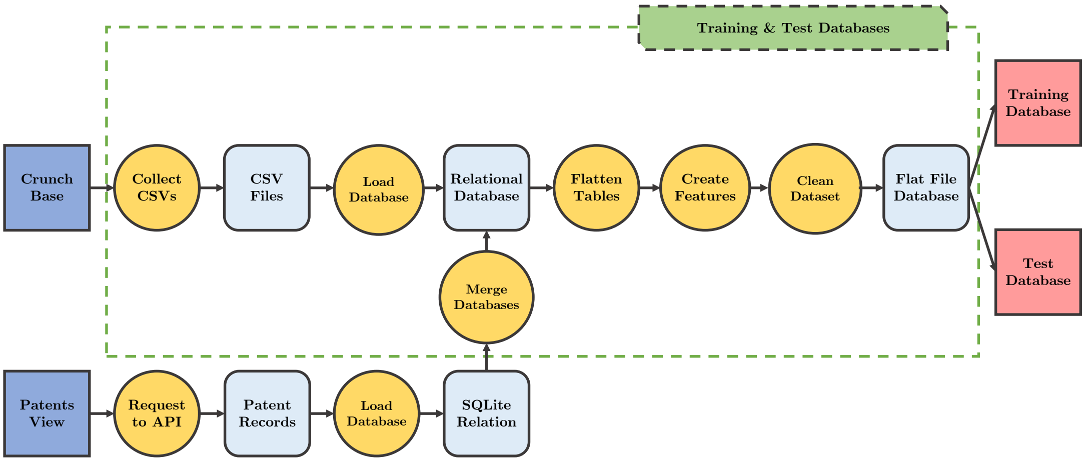


Figure 3.2: Data collection overview.

3.1.1 Data Collection

3.1.1.1 CrunchBase

CrunchBase is an online, crowd-sourced repository of startup companies, individuals and investors with a focus on US high-tech sectors. CrunchBase is freely accessible for browsing but requires licenses to filter the dataset, use the API, and download Microsoft Excel and CSV-formatted dumps. For the purposes of this project, we were granted an Academic License. CrunchBase provides database access in a few formats that offer trade-offs in terms of accessibility and comprehensiveness. We intended to use CrunchBase's API because it provides the most comprehensive access to their database. We developed a collector that downloaded a daily list of updated API endpoints from CrunchBase and queried nodes it needed to update. CrunchBase's API provides JSON-formatted responses which the program recursively parsed and stored into a relational database. However, due to the time constraints of this research project, we abandoned this data collection method. CrunchBase also provides CSV-formatted dumps of their key endpoints (e.g. organizations, people, funding rounds). We downloaded two CSV-formatted dumps from Crunchbase on 09 September 2016 and 04 April 2017 which we loaded into relational databases (see Appendix B for the full database schema).

3.1.1.2 PatentsView

In 2015, the United States Patent and Trademark 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 [40]. The database provides comprehensive information on patents, their inventors, their organisations, and locations. We collected the patent filing records of each company in the primary database, focusing on information relating to dates, citations, and patent types. We matched the data sources on standardised company names (removing common suffixes, punctuation etc.) and using normalised Levenshtein distances. 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 master and test databases.

3.1.2 Dataset Manipulation

To prepare the dataset for machine learning, we first flattened the relational database into a single file using SQL aggregation queries. We aggregated each relevant relation in turn, grouping by Company ID and then combined each aggregated table using left outer joins. Following this process, we used Python to convert tuples (e.g. Round Type and Round Date) and lists (e.g. Round Types) into dummy variables.

We performed preliminary screening on the primary dataset (N = 425,934) to ensure it only included relevant companies. We were interested in removing traditional, non-startup businesses from the dataset (e.g. consulting firms, companies that will not take 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 are associated with shifts in a startup company's functions and objectives and we also expect them to correlate with company age. Our dataset as grouped by startup developmental stage is depicted in Figure 3.3.

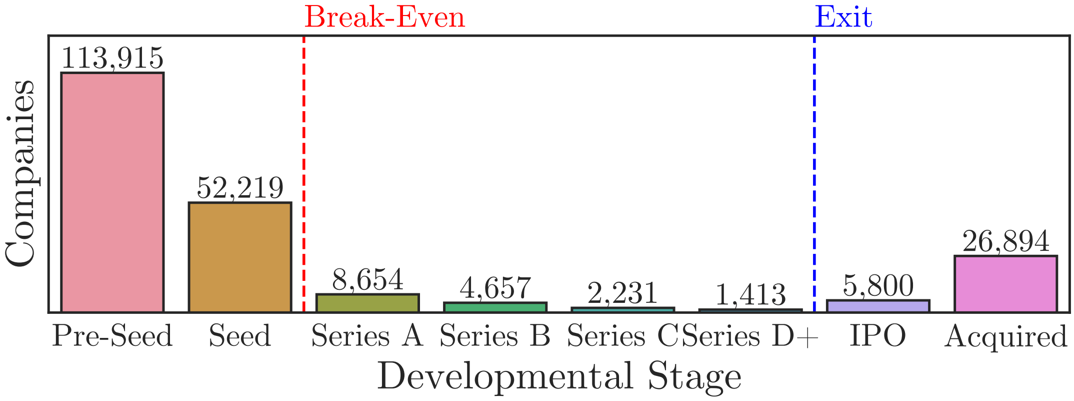


Figure 3.3: Idealised startup development lifecycle with company counts from the master dataset (c. September 2016).

After attempting to place the companies into development stages we are left with a large group of companies (the majority of the dataset) that have not raised funding and so can not be classified on that basis. We assume that companies that have not raised funding fall into two groups - those that intend to raise funding but have not had time to yet, and those that have chosen not to pursue funding and are unlikely to do so. We separated these groups by applying a cutoff 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 theoretically seek investment, we also excluded Closed, Acquired and IPO groups from further analyses (N = 35,973, 8.4%).

Figure3.4 depicts the ages of companies in the master dataset, grouped by developmental stage. While there is significant variability in age for each developmental stage, there is a broad 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 10 years old and the 75th percentile is at 15 years old. On this basis, we excluded companies that are over the 75th percentile of the age of companies in the Series D+ category (N =9,756, 2.2%). Overall, our preliminary screening steps reduced the dataset from 425,934 companies to 153,043 companies, a reduction of 64.1%.

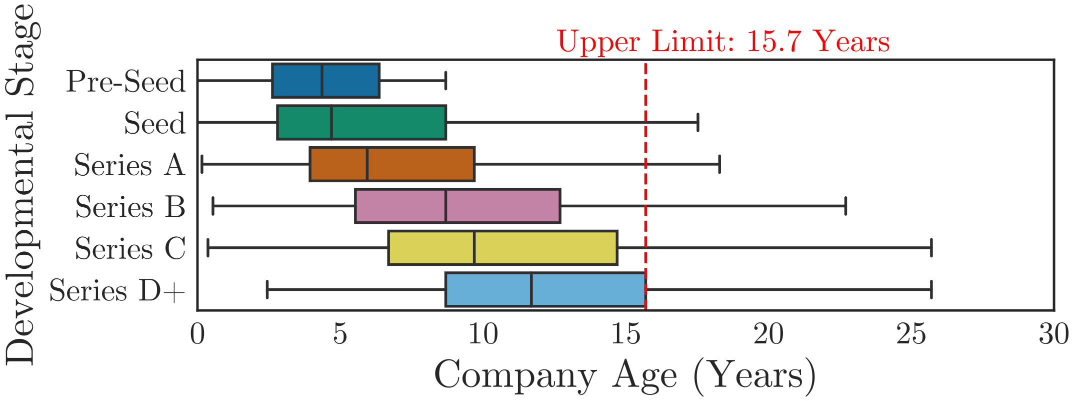


Figure 3.4: Company ages in years grouped by developmental stage. The dashed red line represents the 75th percentile of the age of companies in the Series D+ category (15.7 years).

3.1.3 Exploratory Analysis

3.1.3.1 Descriptive Statistics

Table 3.1 presents the descriptive statistics for the dataset. The dataset is heavily skewed towards New companies (i.e. companies that were recently founded and have not raised any type of funding yet, 68.9%). These companies have very few available features in comparison to companies at later developmental stages. We will investigate the impact of this sparsity on our predictions in Chapter 4. We are presented with a fairly heterogenous dataset, the interquartile ranges imply significant variability in all measures. We do not believe that this implies that the data has not been cleaned effectively, but rather, reflects that startup companies naturally vary significantly in their traits.

CrunchBase's approach to industry classification is simplistic compared to other classification schemes (e.g., USSIC, NAICS, VentureSource) which generally have an industry hierarchy with tiers for broad industry sectors and subsectors providing further granularity. As a result, CrunchBase class labels include over represented and vague classes (e.g., "Software", "Internet Services") which could describe the majority of companies included in the database. In fact, "Software" and "Internet Services" account for 16.4% and 13.4% of all companies in the dataset respectively (see Figure 3.5). Despite these vague class labels, it is

	Obs	Age (Years)		Funding Raised (USD, millions)		Funding Rounds
	N	IQR	Median	IQR	Median	
Pre-Seed	113915	3.737	4.362	0.000	0.000	
Seed	38942	4.003	4.663	1.295	0.250	
Series A	6615	5.005	5.693	7.906	4.400	
Series B	3342	5.918	7.608	22.032	14.891	
Series C	1610	5.523	8.696	45.881	35.285	
Series D+	998	5.005	9.696	90.300	74.385	
Included	165422	4.003	4.688	3.970	0.000	

Table 3.1: Descriptive statistics grouped by developmental stage. Source: Master dataset (c. Sep-16).

clear the dataset skews towards high technology startups, as opposed to biomedical, agricultural, or other technologies.

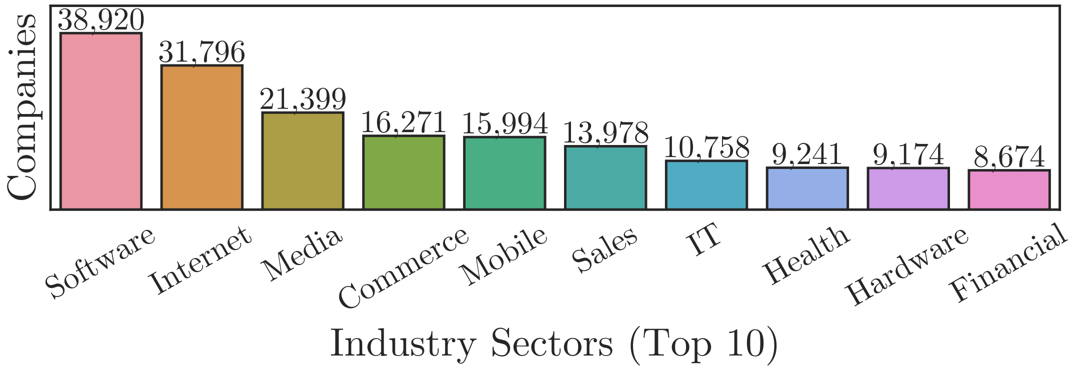
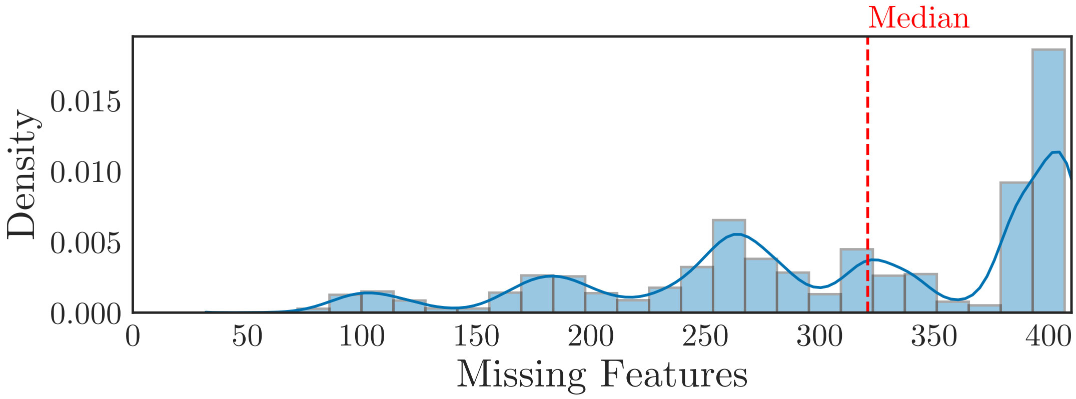


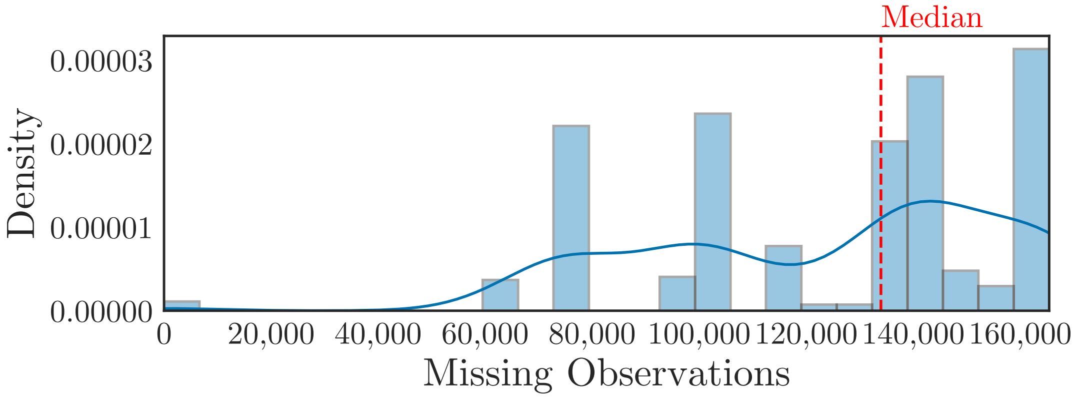
Figure 3.5: Companies grouped by industry sector. The 10 most common sectors are displayed. Source: Master dataset (c. September 2016).

3.1.3.2 Sparsity

First, we explored missing data in the dataset. We expected the dataset to be highly sparse because it primarily came from CrunchBase, a crowd-sourced database. As profiles are entered into CrunchBase piece-meal, it is not clear at face-value whether data (e.g. records of funding rounds) is missing or didn't occur. Figure 3.6 displays the distribution of missing data in the dataset, with respect to each feature and each feature. The multi-modal peaks of both distributions suggest that missing data across certain groups of features may be correlated with each other (e.g. all features derived from funding rounds).



(a) Distribution of missing data by company

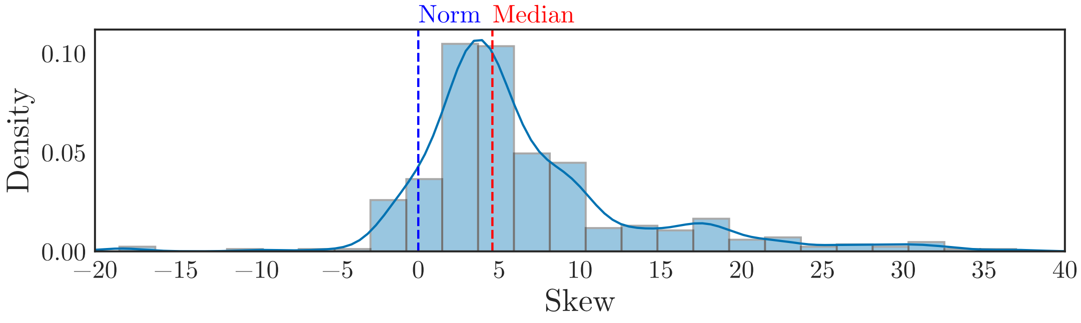


(b) Distribution of missing data by feature

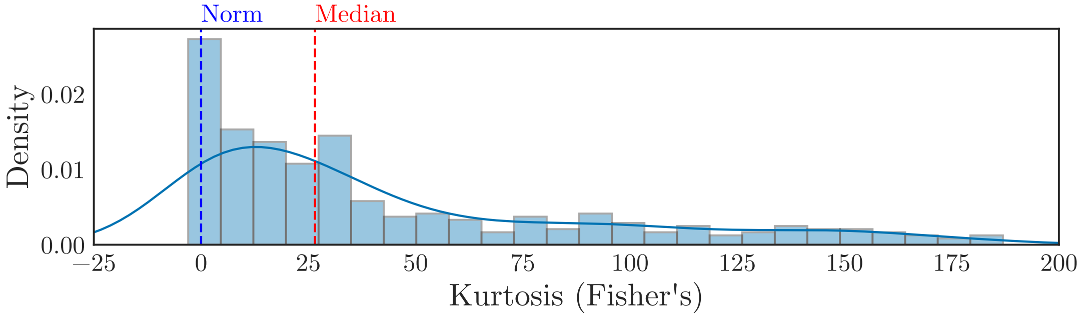
Figure 3.6: Distribution of missing data in master dataset (c. September 2016).

3.1.3.3 Normality

Next, we explored the distributions of features. Figure 3.7 shows the skewness and kurtosis of the features in our dataset. A feature is considered horizontally symmetrical if it has a skewness of 0 and generally considered highly skewed if its absolute skewness is above 1 [9]. The vast majority of our features are more skewed than this cutoff. Kurtosis is a measure of the distribution of variance in a feature. We use Fisher's measure of kurtosis, which has a normal value of 0. Our dataset has consistently higher kurtosis than normal which suggests that we have many extreme values in our dataset. These results in concert suggest that our dataset has many features that are positively skewed with long-tail distributions. This is intuitively what we might expect for features like "amount of funding raised".



(a)



(b)

Figure 3.7: Distribution of features in master dataset (c. September 2016).

3.1.3.4 Scale

Next, we explored the scaling and range of each of our features. Figure 3.8 shows the Interquartile Range (IQR) of each feature (transformed by log1p for ease of viewing). The distribution is extremely skewed, which shows that our features have very different scaling. This may be an issue for our machine learning estimators and feature extractors, so we address this by applying a scaler in our pipeline.

3.1.3.5 Orthogonality

Finally, we explored the orthogonality of our features: the extent to which the variance of our features is unrelated. This is a less straight-forward measure. We explore pair-wise inter-correlations between our features and evaluate how many of the inter-correlations are above a particular correlation cutoff, as depicted in Figure 3.9. We use two correlation metrics: Pearson and Spearman. Pearson is more commonly used but Spearman is a ranked metric and may more accurately

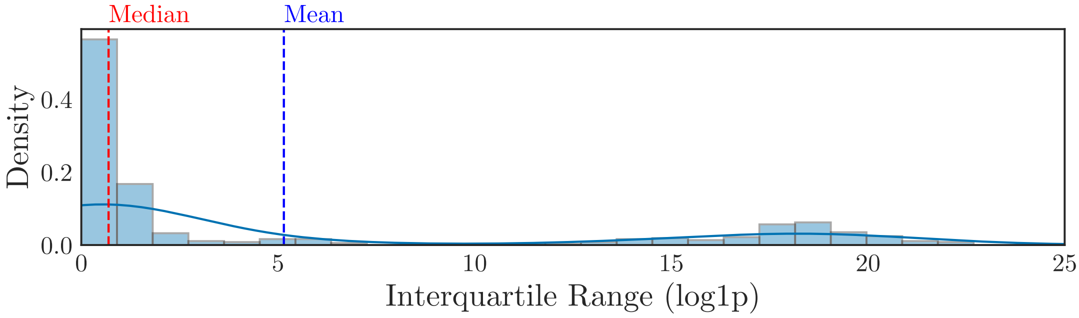


Figure 3.8: Distribution of interquartile ranges (transformed by log1p) in master dataset (c. September 2016).

reflect our non-normal feature distributions. Although most features have relatively low inter-correlations ($\tilde{6}0\%$ below 0.2) there are still a considerable number that are highly correlated, so it might be efficient to remove these features using an unsupervised feature extracter prior to estimation.

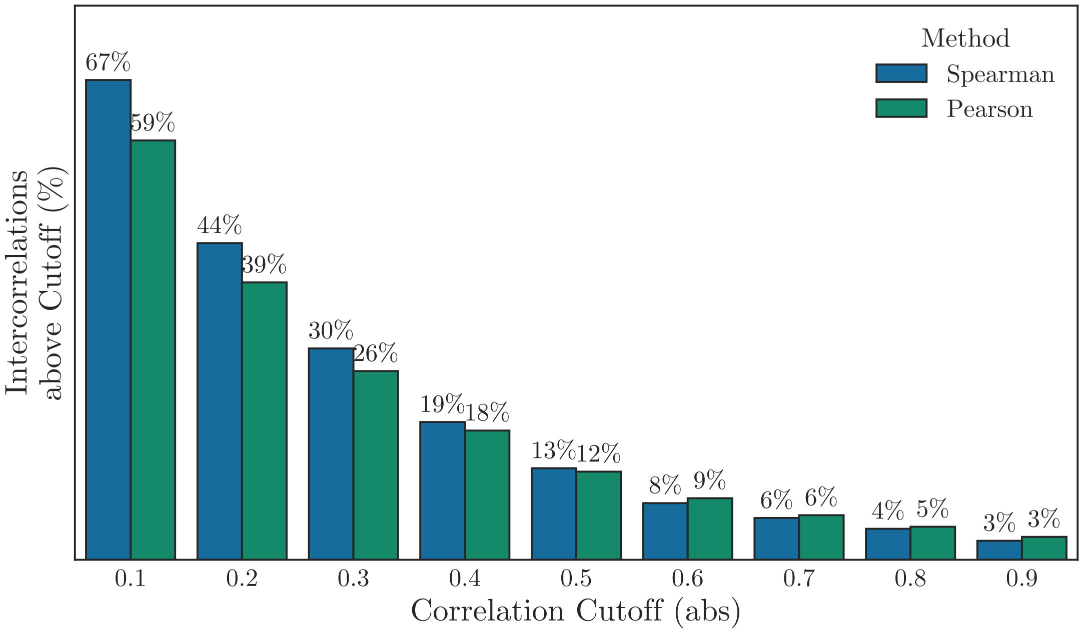


Figure 3.9: Distribution of intercorrelations in master dataset (c. September 2016).

3.2 Pipeline Creation

We developed a classification pipeline using the popular Python-based machine learning library Scikit-learn [34]. The classification pipeline construct allows us to easily search across hyperparameters at each step in the pipeline (see Appendix C for hyperparameter list). The following sections explore the testing of each hyperparameter decision, and the selection of primary classifiers for the following steps. This process is depicted in Figure 3.10.

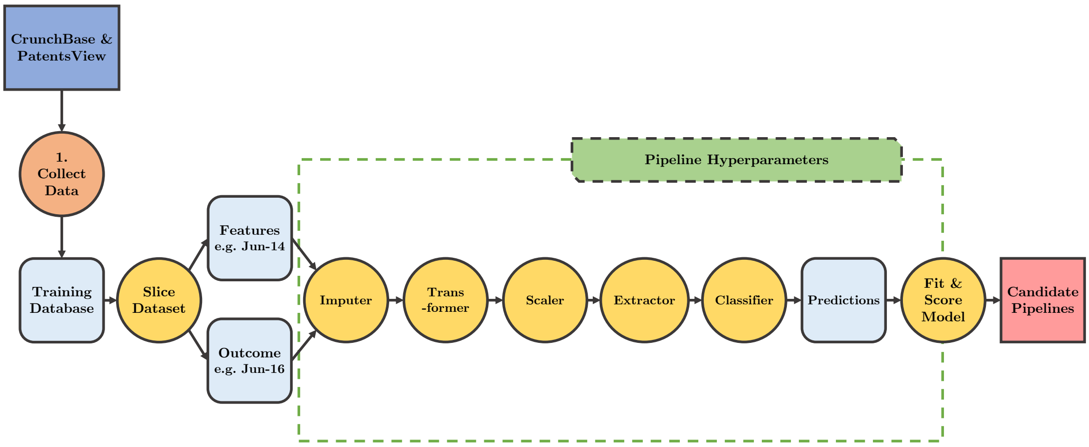


Figure 3.10: Pipeline creation overview.

3.2.1 Imputation

After reviewing the distribution of missing data, we decided to perform further investigation into imputation methods. Common imputation strategies include replacing missing values with the mean, median or mode of each feature. Figure 3.11 shows the distribution of mean, median and modes for each feature in the dataset. For the majority of features, all three measures of central tendency are equal to zero. This resolves the issue of distinguishing missing data from negative observations because, following imputation, all of these data points will map to zero. Figure 3.12 shows the receiver-operating characteristics of the different imputation strategies. As expected, all three imputation strategies produce similar results (within the margin of error).

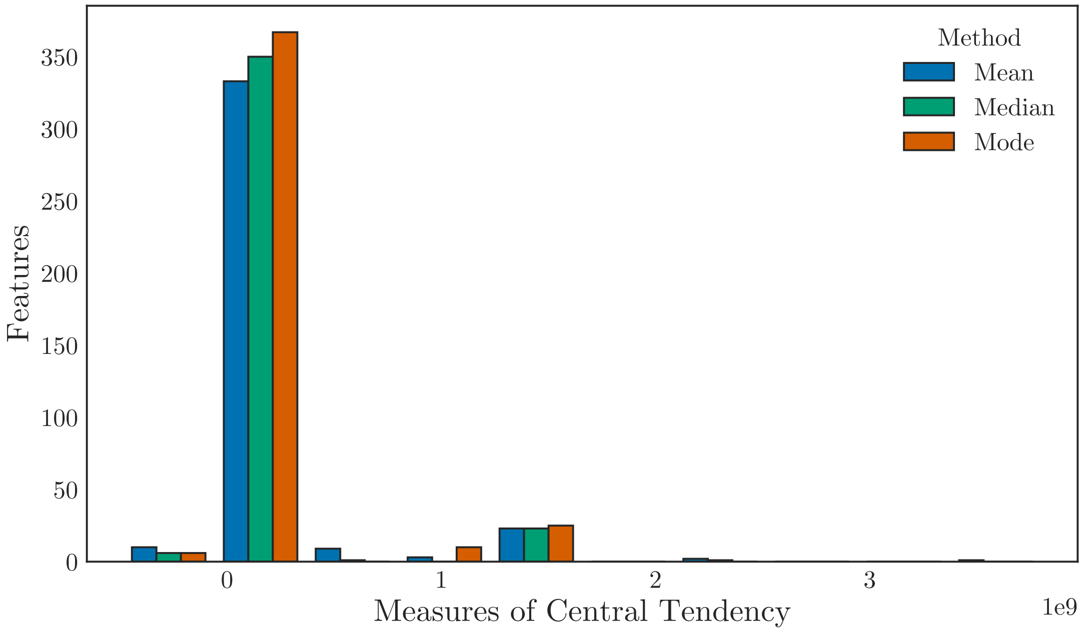


Figure 3.11: Distribution of measures of central tendency (mean, median and mode) in master dataset (c. September 2016).

3.2.2 Transformation

While the classification algorithms we identified in the previous chapter are relatively robust to violations of normality, it may be beneficial to transform the data if the feature distributions are extreme. Table 3.13 shows one of the key features, Total Funding Raised, under a number of different transformations. Like many features in our dataset, the distribution of Total Funding Raised is highly skewed. The log transformation reduces this skew (a normal distribution of non-zero values is apparent) and square root transformation also reduces this skew (to a lesser extent). The impact of these transformations is reduced by the extent of their zero-inflation. However, it is still reasonable to expect both of these transformations to improve the classification accuracy. Figure 3.14 shows the ROC of these different transformation functions. Both functions provide a small performance improvement, with the square root function narrowly best.

3.2.3 Scaling

Standardisation of datasets is a common requirement for many feature extraction methods and machine learning estimators. Sci-kit learn provides three pri-

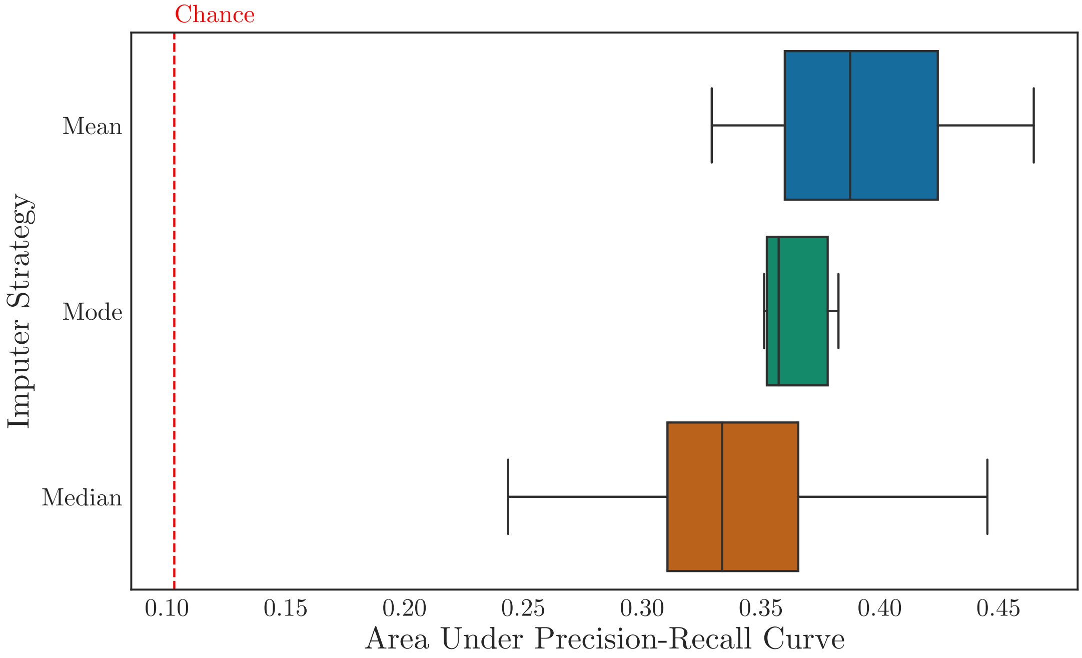


Figure 3.12: Area under Receiver Operating Characteristic (ROC) for different imputation strategies. Imputation strategies include replacing missing values with the most frequent (mode), median and mean value of each respective feature. Results presented are aggregated from hyperparameter optimisation performed over entire classification pipeline (including all classifiers).Source: Features (Apr-12) and labels (Apr-14, 2 year forecast window) derived from Master dataset (c. Sep-16).

mary scaling functions: StandardScaler, RobustScaler and MinMaxScaler. RobustScaler is intended to alleviate the effect of outliers while MinMaxScaler is intended to preserve zero entries in sparse data - both of these are relevant properties for the dataset. Figure 3.15 shows the receiver-operating characteristics of the different scaling functions. MinMaxScaler and RobustScaler actually underperform the null condition while StandardScaler only performs on par with the null condition. This is unexpected but may be caused by the previously applied transformations.

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

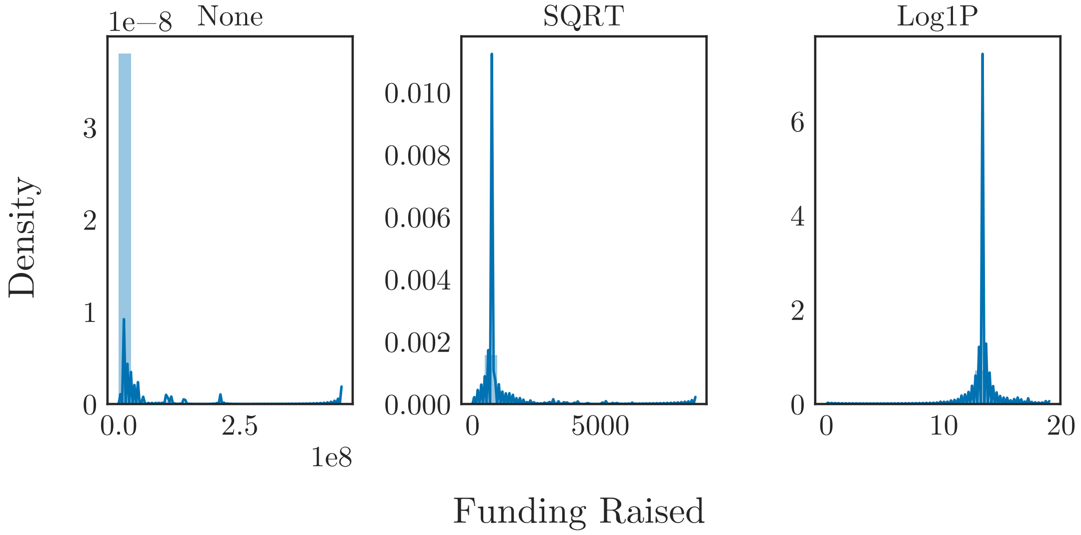


Figure 3.13

approach to dimensionality reduction is Principal Component Analysis (PCA), which constructs orthogonal eigenvectors (components). The magnitude of each eigenvector (its eigenvalue) is displayed in Figure 3.16. The majority of explained variance is captured in the first 10 components, and the Eigenvalues drop below 1 by 100 components - this suggests that these are reasonable values for further hyperparameter search. Figure 3.17 shows the ROC for different numbers of extracted components. All curves produce similar classification results (within margin of error) which implies that we should extract between $1 - 20$ components because it will provide us with more efficient computation.

While PCA is efficient at reducing features, the resultant components are not interpretable. Similarly, individual analysis of 400+ features is difficult to interpret. A compromise is to group the features using the conceptual framework we developed earlier from the literature review. The grouping approach applied weights to each individual feature that optimised the inter-correlations within each group. Given the highly skewed features, we use Spearman correlation which is robust to skewness because it is based on ranking. Figure 3.18 displays the inter-correlations between each factor from the proposed conceptual framework. As we would expect, ‘investors’ and ‘funding’ features are highly correlated. While ‘investors’ attempts to capture the influence of previous investors, it also captures features like the size of an investor’s past investments, which would likely correlate with the size of the investment they made in the target company. Interestingly, ‘founders’ features are positively correlated with all other features except for ‘advisors’ features which are negatively correlated

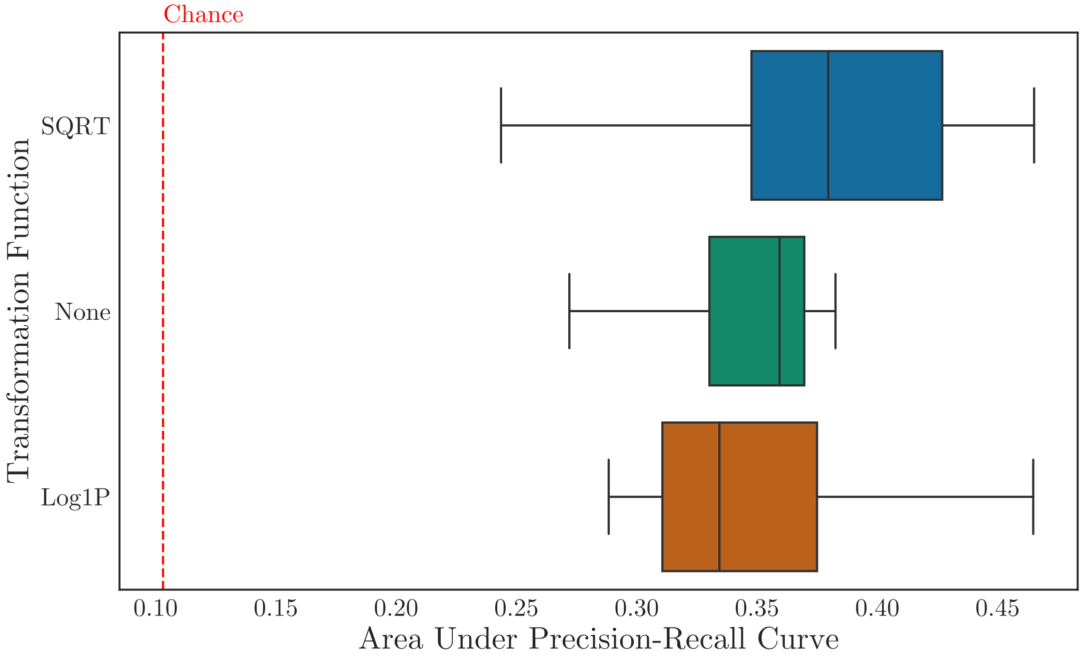


Figure 3.14: Area under ROC for different transformation functions. Transformations include: None (identity transformation), Log1p (natural logarithm of one plus the input array, element-wise), and Sqrt (the square root of the input array, element-wise). Results presented are aggregated from hyperparameter optimisation performed over entire classification pipeline (including all classifiers).Source: Features (Apr-12) and labels (Apr-14, 2 year forecast window) derived from Master dataset (c. Sep-16).

with all other feature groups.

3.2.5 Classification Algorithms

The literature review we performed in the previous chapter revealed seven common supervised classification algorithms potentially suitable for application to this problem area. Our review suggested that Random Forests were most likely to provide a successful 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 as well as median recorded scores to ensure we didn't penalise algorithms that had unfavourable hyperparameter search spaces.

We take a closer look at the Precision-Recall (PR) curves for each classifier in

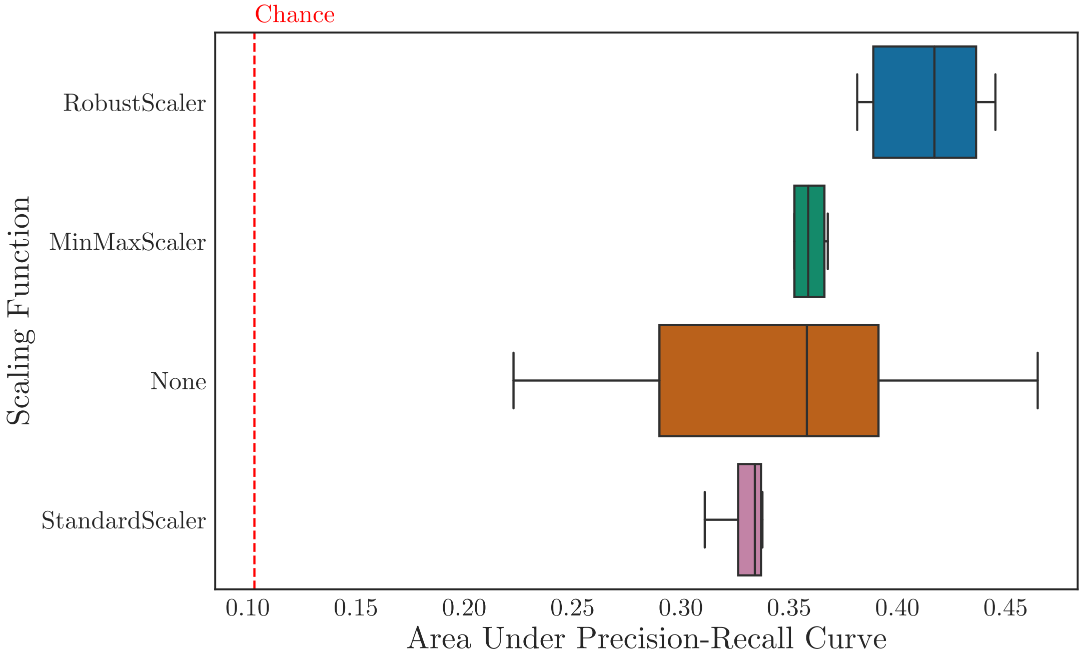


Figure 3.15: Area under ROC for different scaling functions. Scaling functions include: None, StandardScaler (mean: 0, variance: 1), RobustScaler (median: 0, IQR: 1) and MinMaxScaler (min: 0, max: 1). Results presented are aggregated from hyperparameter optimisation performed over entire classification pipeline (including all classifiers).Source: Features (Apr-12) and labels (Apr-14, 2 year forecast window) derived from Master dataset (c. Sep-16).

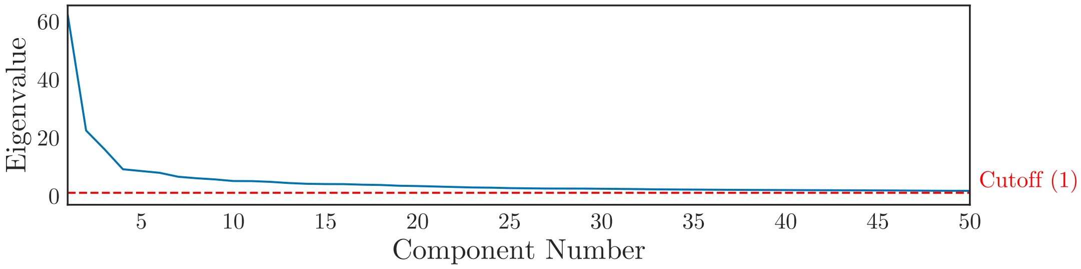


Figure 3.16: Eigenvalues extracted from PCA model. Horizontal line drawn at an Eigenvalue of 1 – this theoretically represents the contribution of one original feature and is commonly used as an approximate threshold for included components. Source: Master dataset (c. Sep-2016).

Figure 3.19. While all classifiers perform better than chance, Logistic Regressions and Random Forests come out ahead, and Support Vector Machines and Artifi-

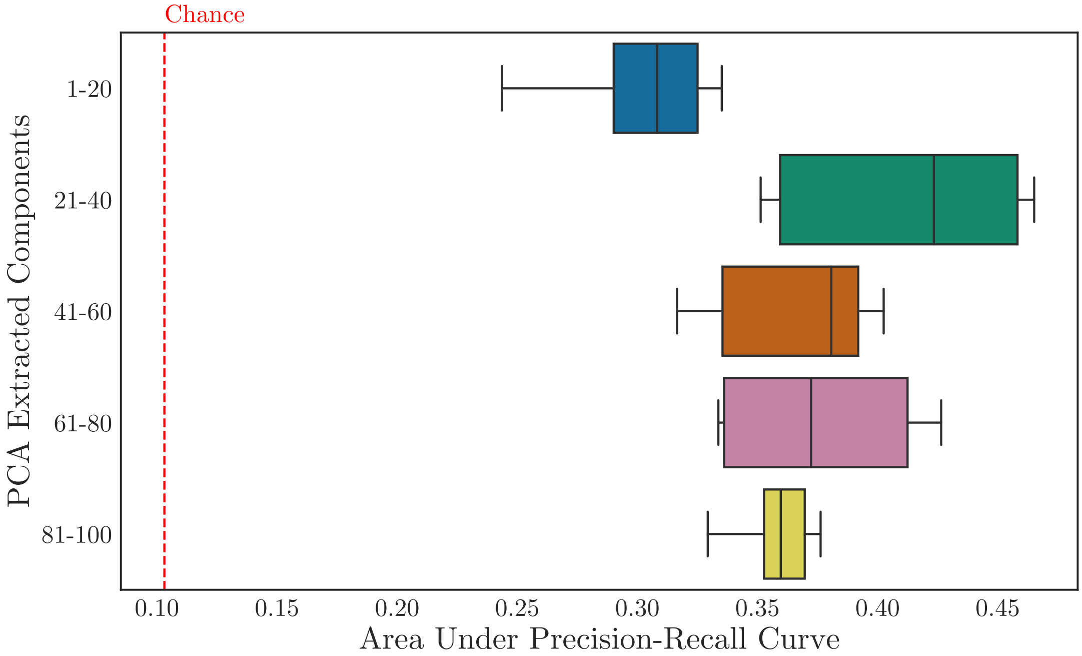


Figure 3.17: Area under ROC for different number of extracted components from PCA. Curves have been grouped by the quotient of the number of components divided by 20 to result in five ordered groups (e.g. Range [0, 19] −¿ 0). Results presented are aggregated from hyperparameter optimisation performed over entire classification pipeline (including all classifiers).Source: Features (Apr-12) and labels (Apr-14, 2 year forecast window) derived from Master dataset (c. Sep-16).

	AUC PRC		AUC ROC		F1		Fit Time (s)	
Classifier	Median	Max	Median	Max	Median	Max	Median	IQR
Logistic Regression	0.417	0.465	0.675	0.710	0.339	0.358	7.326	407.031
Random Forest	0.376	0.465	0.619	0.709	0.332	0.360	68.325	29.064
Decision Tree	0.388	0.429	0.651	0.659	0.305	0.314	15.329	6.747
Naive Bayes	0.354	0.367	0.623	0.638	0.303	0.321	8.589	19.131
K-Nearest Neighbors	0.335	0.353	0.532	0.565	0.131	0.226	8.537	12.316
Artificial Neural Network	0.320	0.335	0.517	0.523	0.072	0.096	9.146	12.017
Support Vector Machine	0.233	0.244	0.503	0.504	0.014	0.017	29.015	0.000

Table 3.2

cial Neural Networks appear to underperform. Delving into the cross-validated learning curves for each classifier (Figure 3.20) we see that Naive Bayes, Logistic Regression, Artificial Neural Networks and Support Vector Machines quickly converge, whereas Decision Trees, Random Forests and K-Nearest Neighbors require more observations to converge. This suggests that we might expect Random

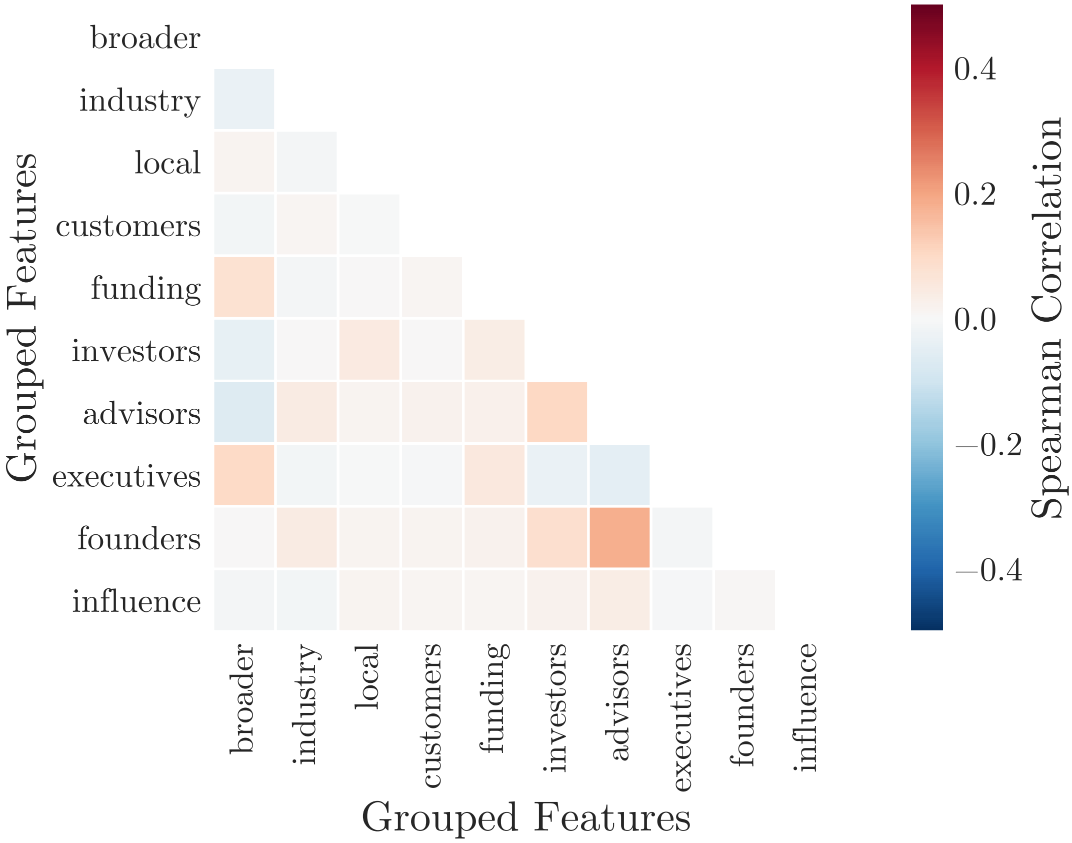


Figure 3.18: Inter-correlations of each factor from conceptual framework. Spearman ranking correlation is used. Individual features are grouped by applying weights that maximise the inter-correlations within each group from our conceptual framework (see Figure 2.1). Source: Master dataset (c. Sep-2016).

Forests to do better in final testing (as testing will not be cross-validated), as well as in the future as the dataset naturally grows.

3.3 Pipeline Selection

In the previous chapter, we developed a system that generated a cross-section of candidate pipelines with different hyperparameters. In this step, we terank these candidate pipelines and evaluate the best pipelines (finalist pipelines) over a number of different dataset slices. This process, depicted in Figure 3.21, ensures that our final pipeline is robust in their performance with respect to time. We aggregate the results for each finalist pipeline across these dataset slices and

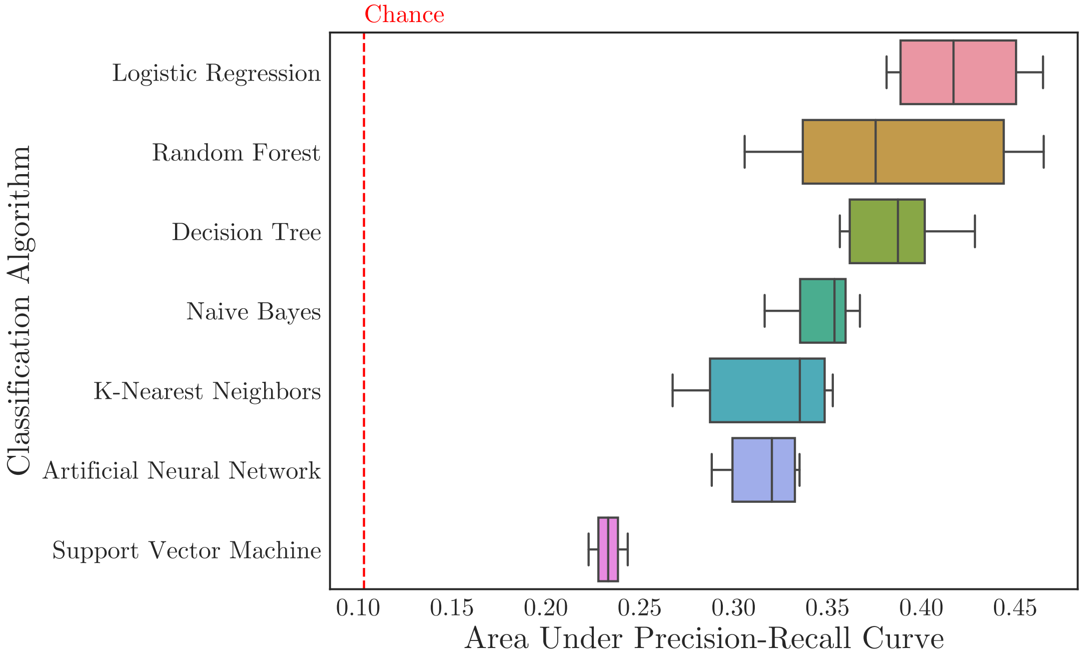


Figure 3.19: Area under ROC for different classification algorithms. All algorithms are implementations from the Sci-kit learn library. Results presented are aggregated from hyperparameter optimisation performed over entire classification pipeline (including all classifiers).Source: Features (Apr-12) and labels (Apr-14, 2 year forecast window) derived from Master dataset (c. Sep-16).

rank the finalist pipelines on their overall performance. Finally, we select the best pipeline.

3.3.1 Dataset Slicing

We developed a procedure for generating historical datasets from our CrunchBase and PatentsView data. CrunchBase provides created and last-updated timestamps for each record in their CSV-formatted dumps (and also in the JSON-formatted responses from their API). We took advantage of this to produce a system that reverse-engineers previous database states by filtering the current database by only records that were created by a given 'slice' date.

We performed preliminary testing of our reverse-engineering technique by comparing a historical CrunchBase database collected in December 2013 with a slice from our primary dataset collected in September 2016, as shown in Figure 3.22. While there are some differences, particularly in the IPO counts, we

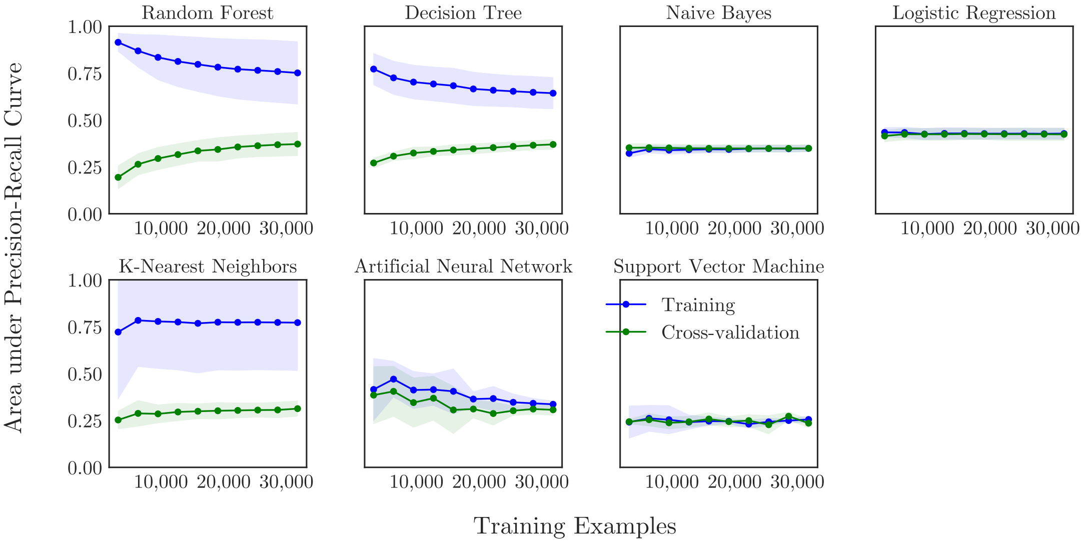


Figure 3.20: Learning curves.

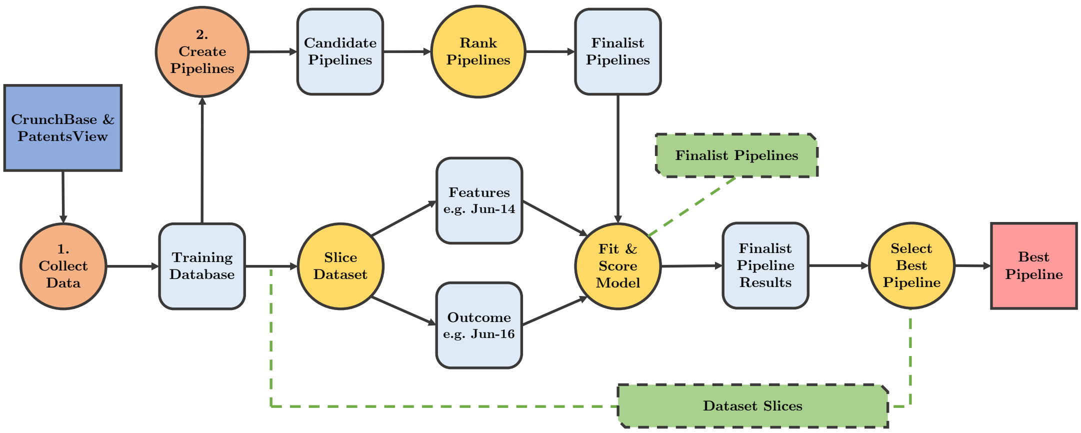


Figure 3.21: Pipeline selection overview.

consider this to be satisfactory variance considering the 3-year time difference (i.e. perhaps some companies have been since removed from the database). The key relations for the purposes of our system are Companies, Funding Rounds and People, all of which had minor differences considering the size of these datasets.

Figure 3.23 shows company counts by startup development stage from different dataset slices. We limited our experiments to dataset slices from 2012-onwards because prior to 2012 the datasets become too small to use to make predictions (particularly given the class imbalance).

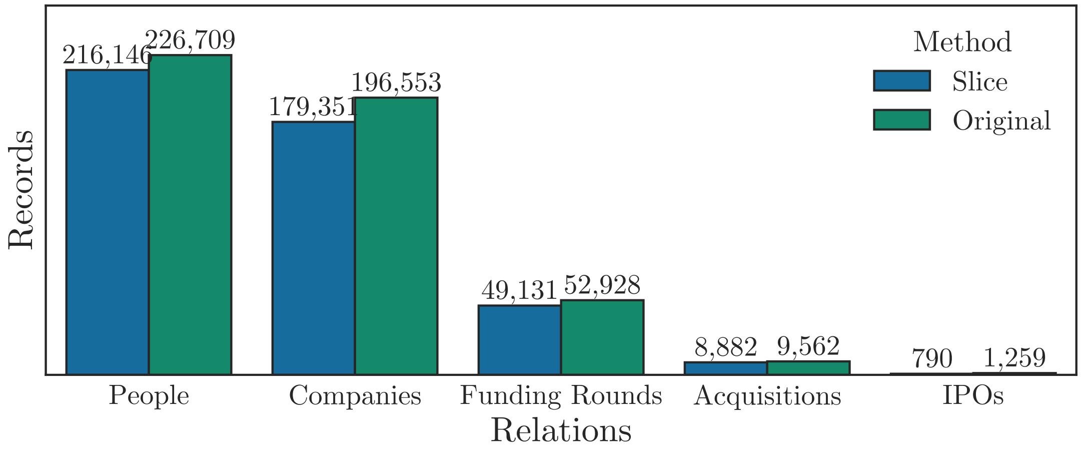


Figure 3.22

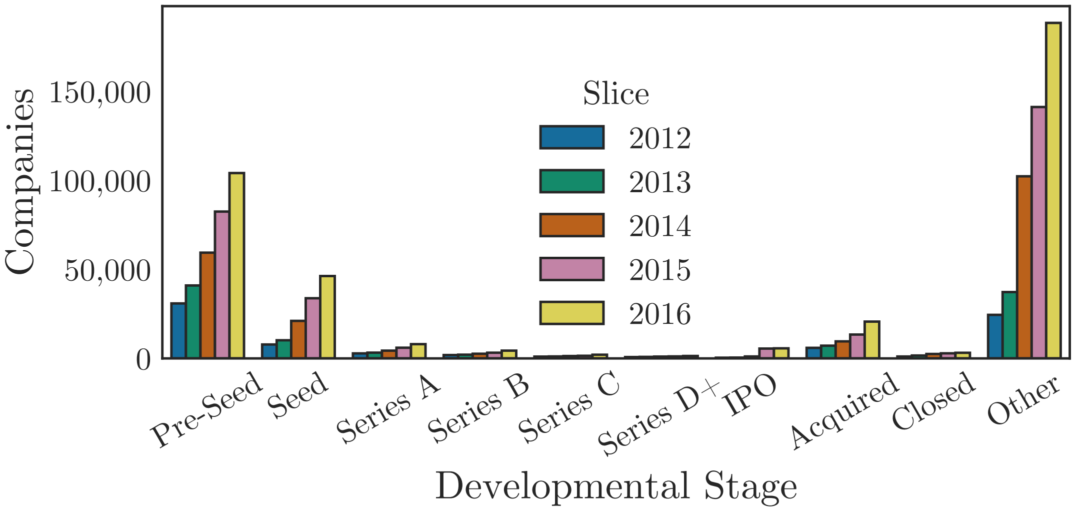


Figure 3.23

3.3.2 Evaluation Metrics

Next, we decided how to narrow our candidate pipelines down to finalist pipelines that we can evaluate further. There are a number of different metrics used to evaluate binary classifiers in machine learning. The most simplistic metric is accuracy but this is rarely used in practice because it gives misleading results in the case of imbalanced classes. Receiver Operating Characteristic (ROC) curves are perhaps the mostly commonly used evaluation tool in binary classification, and 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 accruate results for imbalanced classes than ROC curves [15]. Given our dataset is highly imbalanced (the positive class is approximately 10%) we decided to proceed with PR curves. We will also use this metric to determine which is ultimately the best of our finalist pipelines.

3.3.3 Finalist Pipeline Evaluation

Our hypothesis is that the performance of our classification pipelines may vary with respect to the date that the dataset was collected (in this case, sliced). To study this hypothesis, first we explored variance between the pipelines on aggregate against the slice dates, presented in Figure 3.24. We see little variance on this basis, and we don't observe a relationship between slice date and score.

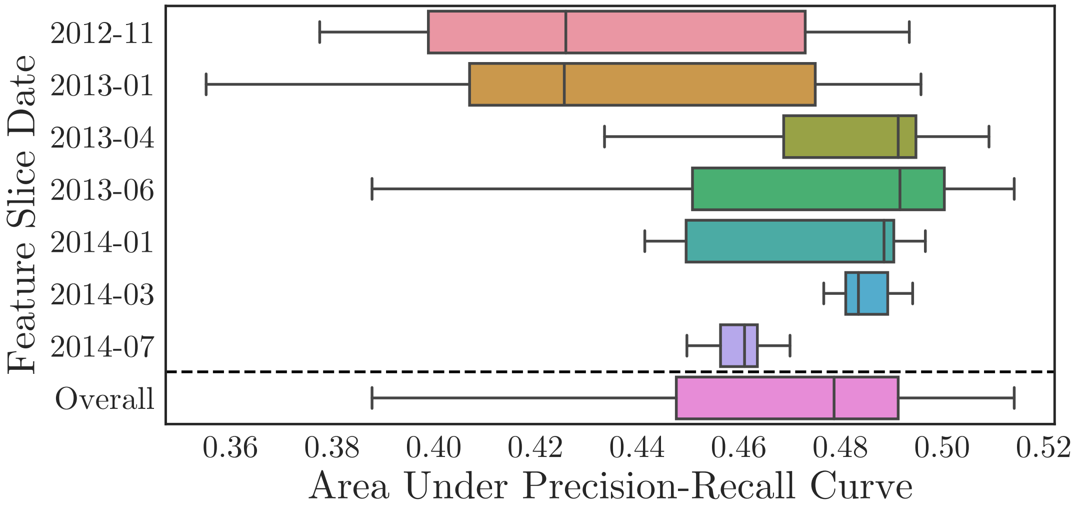


Figure 3.24

Next, we study the variance within the individual pipelines, presented in Figure 3.25. Here, we can see there is significantly more variance in the scores. Although there is still a strong positive correlation between the pipelines initial ranking and their scores, we can see that there are some individual deviations. Importantly, the top-ranked pipeline from the first stage actually has a lower median score than the second-ranked pipeline. These results suggest that the top 3-5 pipelines should be evaluated in this manner to ensure that the best pipeline is selected.

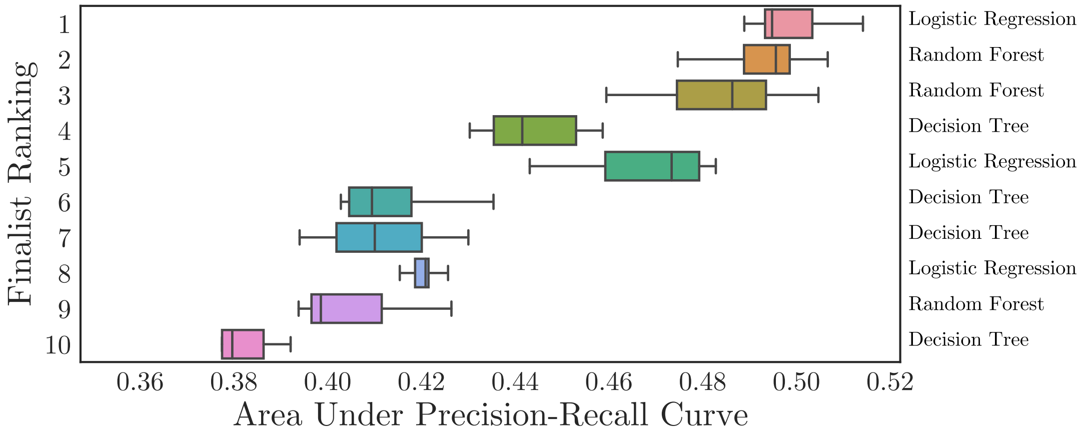


Figure 3.25

The best candidate pipeline is depicted in Table 3.3. We adopted this pipeline configuration for our following experiments.

Step	Parameter	Values
Imputer	Strategy	Mode
Transformer	Function	SQRT
Scaler	Function	MinMaxScaler
Extractor	Components	27
Classifier	Algorithm	Random Forest
Classifier	Class Weight	Balanced
Classifier	Criterion	Entropy
Classifier	Boostrap	True
Classifier	Estimators	34
Classifier	Max Depth	8
Classifier	Max Features	SQRT(N_Features)
Classifier	Min Samples Split	2
Classifier	Min Samples Leaf	1
Classifier	Min Weight Leaf	0
Classifier	Max Leaf Nodes	None
Classifier	Min Impurity Split	1e-7

Table 3.3

3.4 Model Fit and Prediction

Finally, we use our optimised pipeline to produce a model and make predictions, as shown in Figure 3.26. Our system applies the best pipeline that was generated in the previous section to a training dataset, producing a fitted model. The model is then applied to a feature vector from a held-out test database, which generates a set of predictions which could, in practice, then be used by Venture Capital (VC) firms. We evaluate the accuracy of the models produced by our system with respect to a number of variables in the next chapter, Chapter 4.

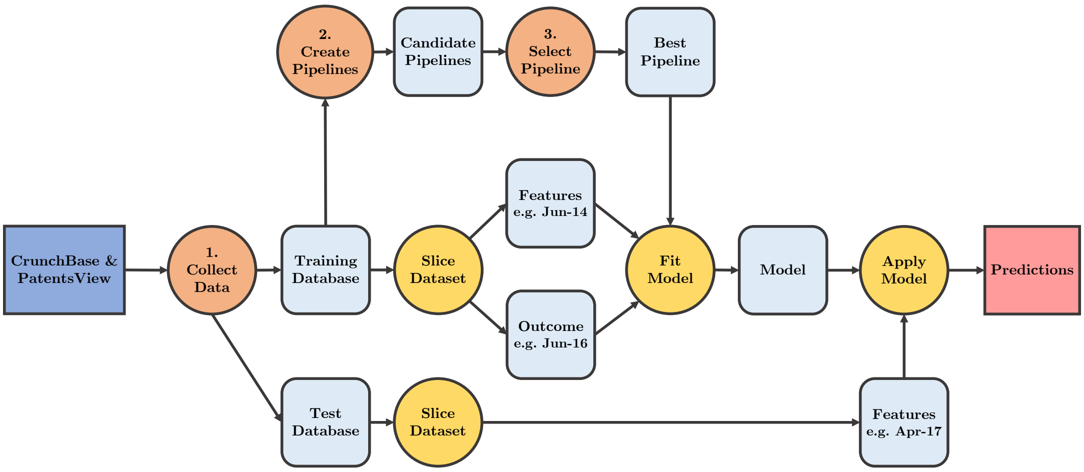


Figure 3.26: Model fit and prediction overview.

CHAPTER 4

Evaluation

We believe it is possible to produce a Venture Capital (VC) investment screening system that is efficient, robust and powerful. In Chapter 3, we described the development and structure of such a system. Our system identifies startup companies likely to receive additional funding or exit in a given forecast window. This system generates statistics and make recommendations that may assist VC firms to efficiently and effectively screen investment candidates. In this chapter, we evaluate models developed by our system against criteria of efficiency, robustness and predictive power.

We produced a classification pipeline optimised with respect to its robustness over time, and evaluated models produced by this pipeline against a held-out test dataset. This evaluation process is depicted in Figure 4.1. The pipeline is fit to a training dataset. The model is applied to a test feature vector to produce predictions. We score these predictions against truth values derived from the held-out test database (collected in April 2017). This process is performed multiple times to evaluate the three primary criteria derived from our literature review: efficiency, robustness and predictive power.

While Area under the Precision-Recall (PR) Curve were used to guide the development of our system during pipeline creation and selection, 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. In this sense, 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're more sensitive to classification performance for the positive class (companies that have been successful in raising further funding or achieving an exit), so thereafter, 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. It should produce similar results to a macro-averaged F1 Score, incorporating the performance of both classes.

Firstly, we evaluated efficiency by exploring the learning curves of our classification techniques and whether there is sufficient data to produce reliable statistics. We also explored the time profile of our system and whether it is reasonable for use in industry, and would be likely to reduce the time currently taken to perform similar analyses. Secondly, we evaluated robustness by evaluating our models against multiple reverse-engineered historical datasets and measuring their variance. Thirdly, we evaluated the system's predictive power across different forecast windows, for startups at different stages of their development lifecycle, and for different potential target outcomes.

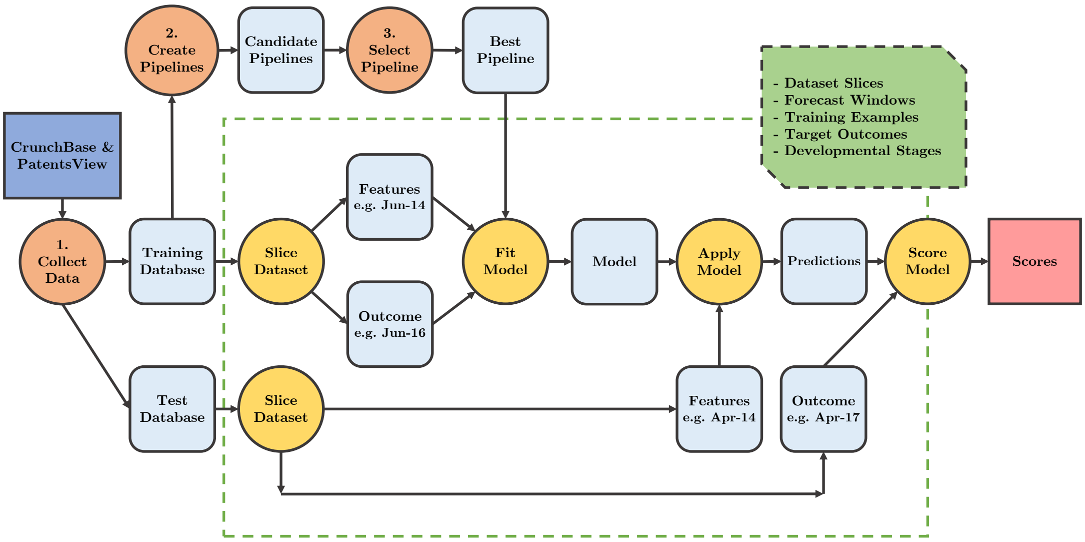


Figure 4.1: Pipeline evaluation overview.

4.1 Efficiency

The Venture Capital (VC) industry requires more efficient forms of investment analysis, particularly in surfacing and screening. These processes are currently performed through referral, Google search, industry papers and manual search of startup databases. By its nature, our automated system should be more efficient than these methods. In this section, we assess how efficient our system is – in terms of data consumed and time taken – and look at whether we can further improve its efficiency.

4.1.1 Dataset Size

Learning curves allow us to evaluate how the bias and variance of a classification technique varies with respect to the amount of training data available. We investigated learning curves for our classification pipeline to determine whether smaller samples could achieve similar predictive power and reduce the system's computational demand. We applied 10-fold stratified cross-validation to split our dataset into 10 subsets of different sizes which we used to train the estimator and produce training and test scores for each subset size. The rate of convergence of our training and cross-validation curves implies whether our classification pipeline is over- or under-fitting our data for various sizes allowing us to select an optimal sample size.

Figure 4.2 shows the learning curves for forecast windows of 2-4 years. 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 4 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 that this pipeline configuration is approaching peak performance. When the system is run in the future (with a larger dataset), the pipeline creation process may choose a classifier with less bias and more variance, like a Support Vector Machine (SVM) or Artificial Neural Network (ANN).

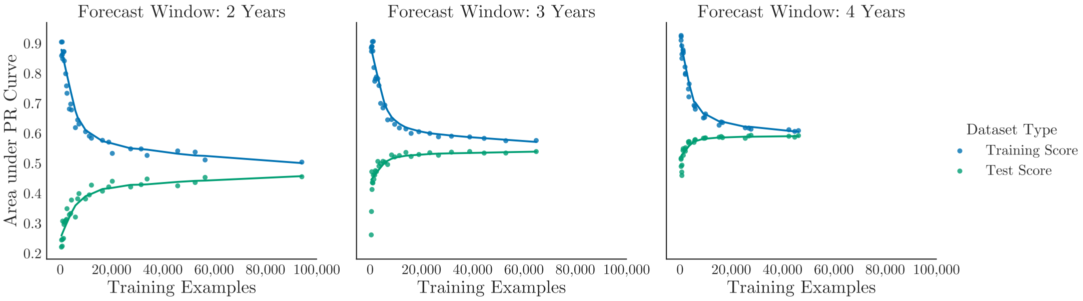


Figure 4.2: Learning curves by forecast window.

The plots in Figure 4.2 are evaluated against our base target outcome, which we term "Extra Stage" (i.e. whether a company raises an additional funding round, is aquired or has an IPO). When our learning curves are split by components of this target outcome, we see that the efficiency of our system varies, as shown in Figure 4.3. We observe that predicting whether a company raises an extra round is the least data-intensive outcome, as it converges rapidly even over a forecast window of 2 years. In comparison, predicting company exits does not converge, even over a forecast window of 4 years. Our model has most difficulty predicting IPO exits, which are rare events even in our large dataset. For these target outcomes, we would expect our system would benefit from a larger dataset. It should be noted, however, that our pipeline was optimised for predicting our base target outcome, and if the entire system was performed on the different target outcomes we might find other classification pipelines provide better performance.

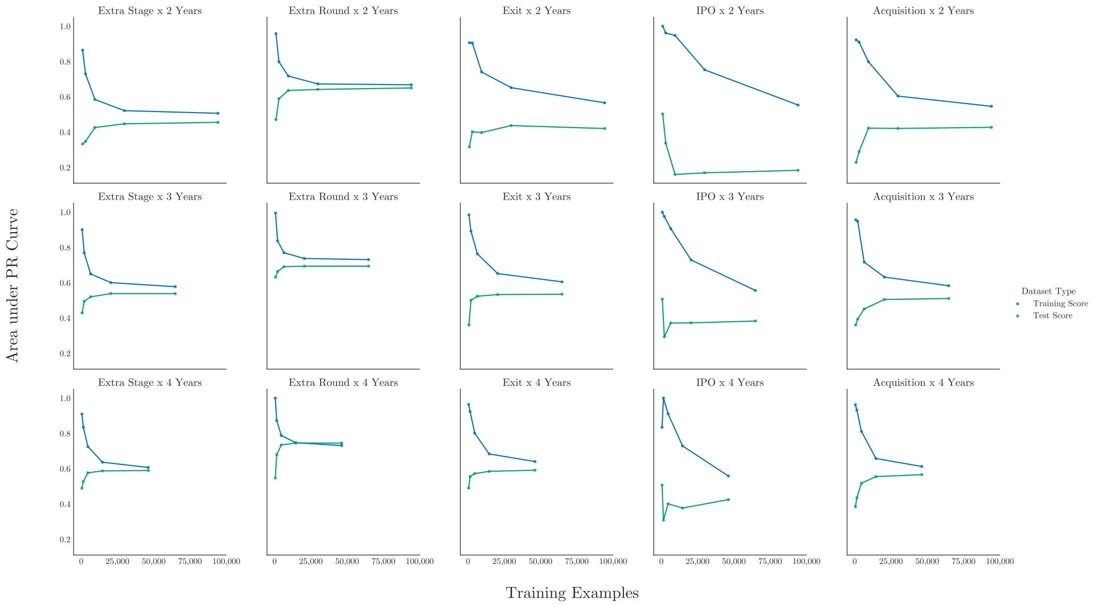


Figure 4.3: Learning curves by target outcome (column) and forecast window (row).

4.1.2 Time Profile

Unlike other forms of finance, like equity or derivatives trading, VC operates on a much longer timeframe – deals typically close over weeks, rather than minutes. This has two key disadvantages: VC firms have higher management costs because they spend more time screening investments and startup founders waste precious time negotiating with investors when they could be building their businesses. Automated systems could significantly decrease the time taken to generate investment opportunities. We investigated 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 approximately 46 hours to complete on a modern desktop PC. When we further break this time down by system component, it's clear that 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 particularly long time because, in this case, it also involves generating learning curves for reporting, which is another cross-validated process. However, when placed into production, this component could be run infrequently - perhaps once per year - to ensure that the pipelines being used are still optimally suited for the dataset. The next component of the system, selecting the most robust pipeline, could occur more frequently - perhaps once every month - and the final component of the pipeline, making up-to-date predictions, could be evaluated every time new data is fed into the system (perhaps once per day) because it only takes an hour.

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.

4.2 Robustness

The Venture Capital (VC) industry is concerned that predictive models trained on historical data will not accurately predict future trends and activity. This has been identified as a key barrier to the adoption of automated systems by the VC industry [46]. 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.

We generated three models from datasets created from our training database from each year of 2012-2014 for forecast windows of 2 years (i.e. [2012, 2014], [2013, 2015], and [2014, 2016]) and evaluated each model against a dataset created from our test database (i.e. [2015, 2017]). We expected that if the factors that predict startup investment success through time are consistent, we would observe little difference between the performance and characteristics of these models.

Figure 4.4 shows the coefficient of variation of models trained on dataset slices from different years, against key evaluation metrics. The coefficient of variance is the ratio of the biased standard deviation to the mean. This produces a standard measure of variance, so different evaluation metrics are comparable. We have also grouped by forecast windows as later dataset slices cannot be tested with long forecast windows which skews results along this dimension. Variance across all metrics is very low, with slightly more variance over shorter forecast windows, as one would expect.

We explored the feature weights for each model in Figure 4.5. While there are some slight differences, the general trend is very similar across all models. We will discuss the distribution of these feature weights in more detail in a following section.

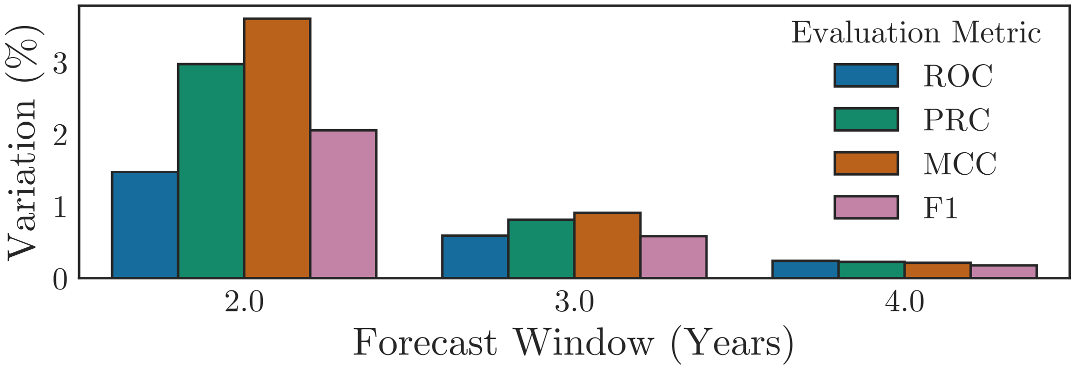


Figure 4.4: Performance variation by slice date.

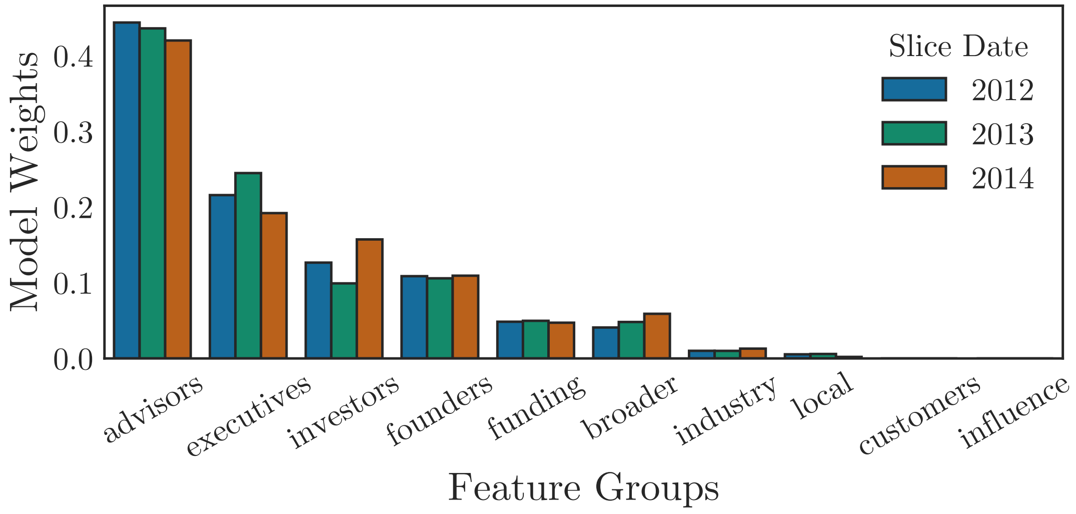


Figure 4.5: Feature weights by slice date.

4.3 Predictive Power

The system must be consistently accurate at identifying a variety of high-potential investment candidates. We evaluated the systems' predictive power 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 IPO, or some combination thereof).

4.3.1 Baseline Analysis

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

First, we looked at 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 difference between the slice that provides our features and the slice that provides our outcome. We combined pair-wise datasets of each year from 2012-2016 inclusive and explored the proportion of companies that raised additional funding or exited.

Figure 4.6 shows how company outcome varies with respect to the forecast window (time between the observed features and the measured outcome). Intuitively, we see a positive relationship between length of forecast window and company outcome. In particular, very few companies appear to have exited or raised funds over a period of less than 2 years so we will focus our experimentation on forecast windows of 2-4 years.

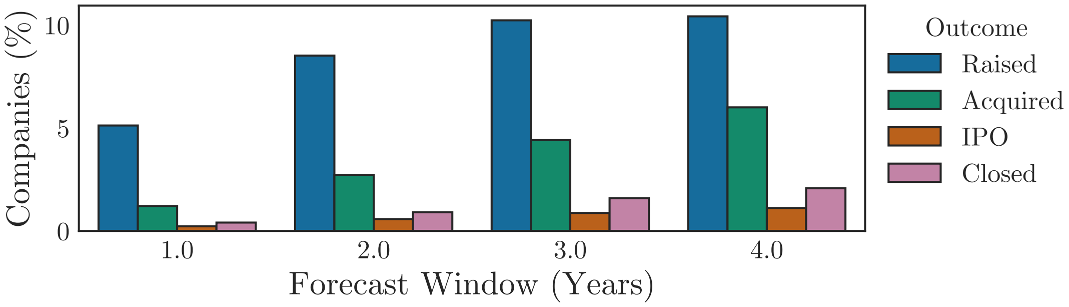


Figure 4.6: Outcomes by forecast window.

We also looked at how company outcomes vary with respect to development stage, shown in Figure 4.7. We see a broad positive relationship between developmental stage and likelihood of further funding rounds and exits, which we would expect as at each stage there is higher market traction and scrutiny from investors. The variance between the outcomes of different developmental stages suggested that in our experimentation we should investigate how our system predicts each stage independently, as well as in aggregate, as we do in a following section.

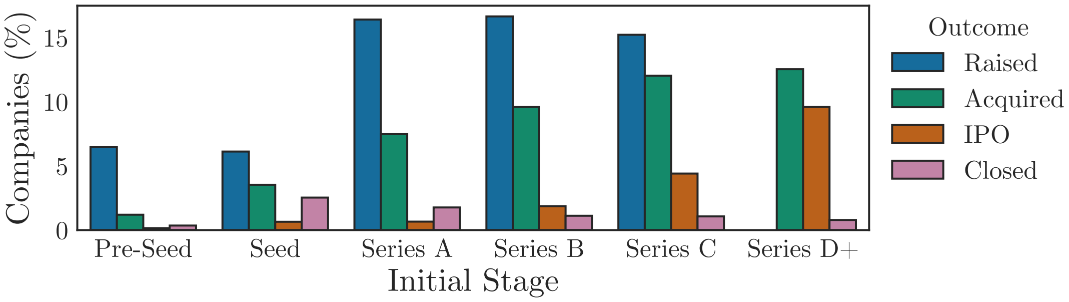


Figure 4.7: Outcomes by developmental stage.

4.3.2 Forecast Windows

A forecast window is the period of time 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 3 years.) The Venture Capital (VC) industry raises funds with fixed investment horizons (generally 3–8 years), 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 varies with respect to the length of these forecast windows.

Figure 4.8 shows model performance across a range of metrics, grouped by forecast window. As discussed previously, we do not expect Area under the Receiver Operating Characteristic (ROC) curve to accurately reflect the performance of our model because it is not sensitive to our bias towards the positive class. We see that there is very little difference in Area under the ROC curve across the forecast windows. However, across all three other metrics, there is a clear positive relationship between length of forecast window and model performance. In particular, 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%), which probably reflects that our F1 Score here purely captures the positive class.

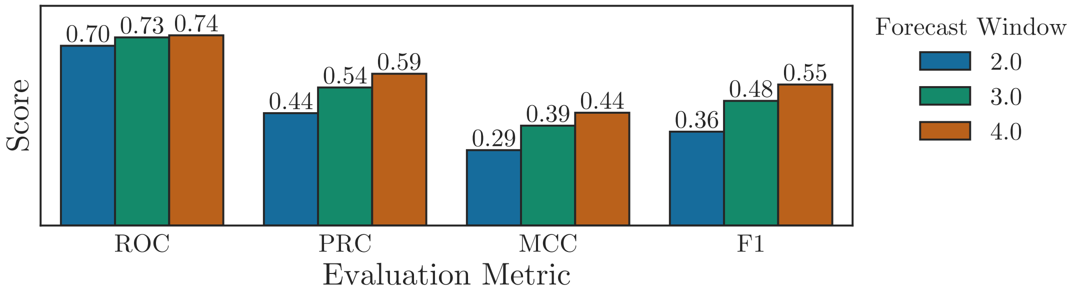


Figure 4.8: Performance by forecast window.

Figure 4.9 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. Advisors, somewhat surprisingly, are the best predictor of startup investment success. This may reflect that exceptional startups are more effective at attracting influential advisors. Executives and founders 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 factors are weak predictors, as are customers and social influence (in this case measured through participation at events). These factors are sparsely represented in the CrunchBase database.

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 advisors increases over time, and the importance of executives and the broader economy decreases over time.

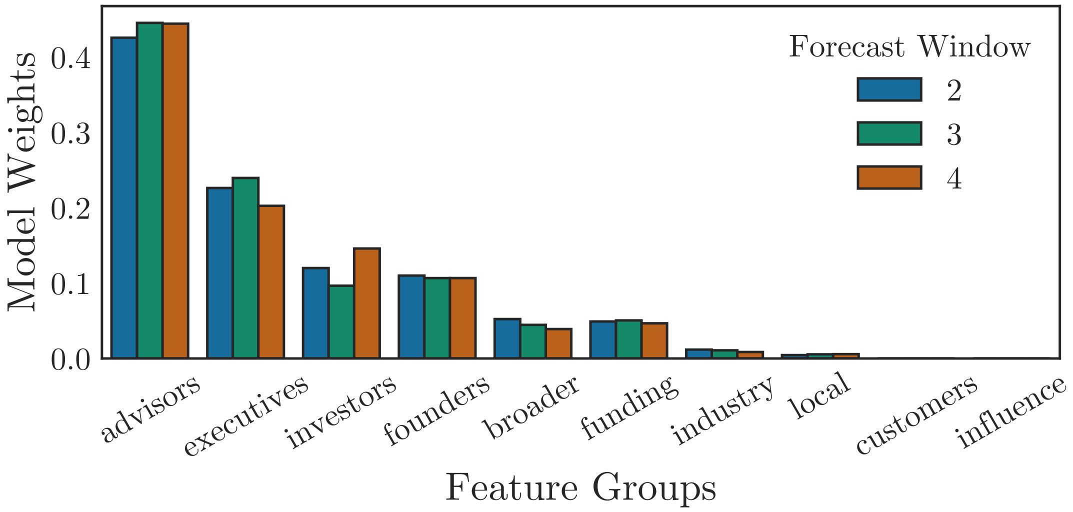


Figure 4.9: Feature weights by forecast window.

4.3.3 Development Stage

Startups can be broadly classified into developmental stages by virtue of their external funding milestones. These milestones not only signal 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 4.10 shows F1 Scores grouped by developmental stage and fit method. First, we exmaine the baseline distribution and then the variation in performance by fit method. Model performance has a positive relationship with developmental stage. This may be a product of later stage companies having more complete feature vectors. The only deviation from this relationship is for Series D+. This may be because the model is only predicting exits at this stage. To understand this discrepancy better, we split the datasets into their developmental stages and fit the model onto each of these sub-datasets individually. This results in a broad performance improvement. Pre-Seed companies make up most of our original dataset and we see the smallest difference between methods for this stage. However, for Series D+ we see a significant difference in performance, which may suggest that the features that predict Series D+ performance are different to in earlier stages.

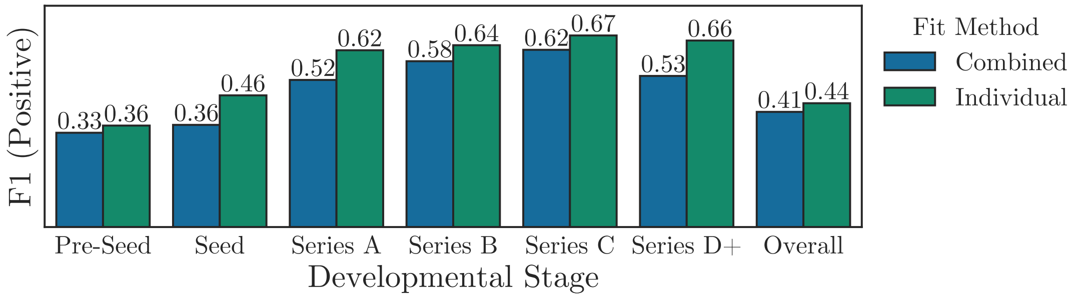


Figure 4.10: Performance by developmental stage.

Figure 4.11 shows the standardised weights of features, grouped by developmental stage. While a similar trend to Figure 4.9 is clear, there is more varation in weights than was observed 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 very important in pre-seed companies, as is the broader economic outlook.

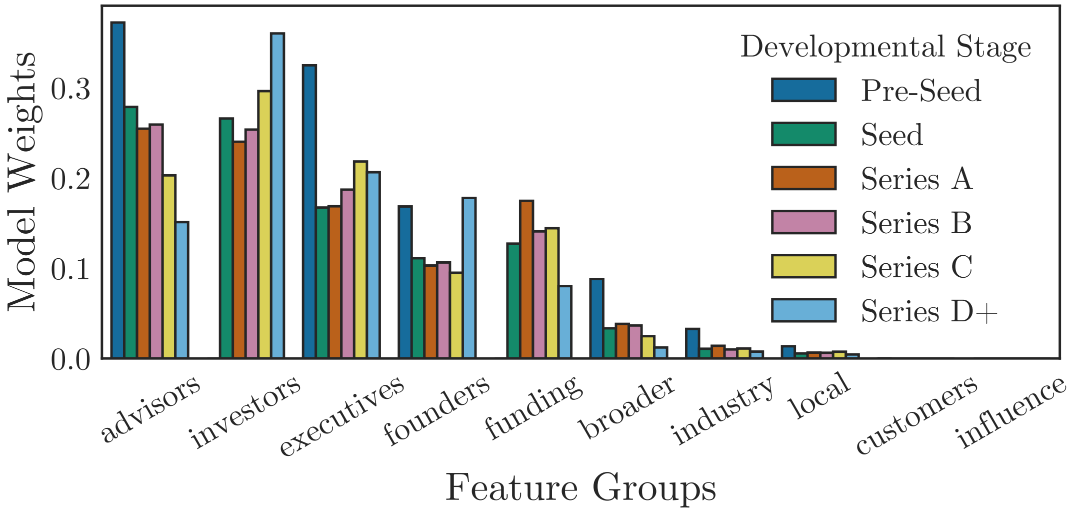


Figure 4.11: Feature weights by developmental stage.

4.3.4 Target Outcomes

Ultimately, VC firms seek rare investments that will return their invested funds many times over within an investment horizon of their fund (typically 3-8 years). Funds are generally only returned to VC investors when startups have liquidity events (IPO, Acquisition). However, particularly recently, many companies that are considered highly successful are delaying their liquidity events (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, which we term "Extra Stage" (i.e. whether a company raises an additional funding round, is aquired or has an IPO). In the following section, we will explore whether the component outcomes (e.g. predicting IPOs) will have an affect on our system's predictive power.

Figure 4.10 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 performs badly at predicting IPOs, though these are rare events in our dataset. As we observed in Figure 4.8, there is a clear positive relationship between length of forecast window and model performance. We observe fairly similar performance improvements across the target outcomes except for in the case of IPOs, which improve dramatically from a forecast window of more than 2 years. This may be because there are non-performance related factors that affect IPO timing, so predicting the timing of an IPO is difficult for our model to achieve.

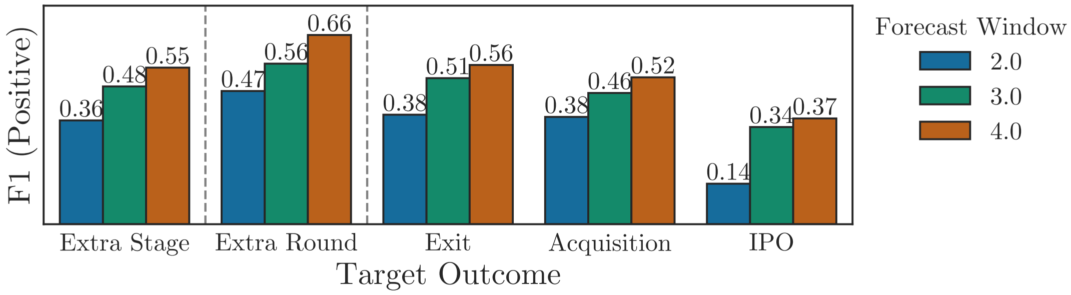


Figure 4.12: Performance by target outcome.

Figure 4.13 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, probably because Acquisitions make up a large proportion of Exits in our database. Investors, Executives and Founders are key features for Exits and Acquisitions. In comparison, IPOs have more weighting towards Funding, Advisors and the Broader Economy. Extra Round is most strongly related to Investors and Funding. Extra Stage is, perhaps surprisingly, most strongly related to Advisors.

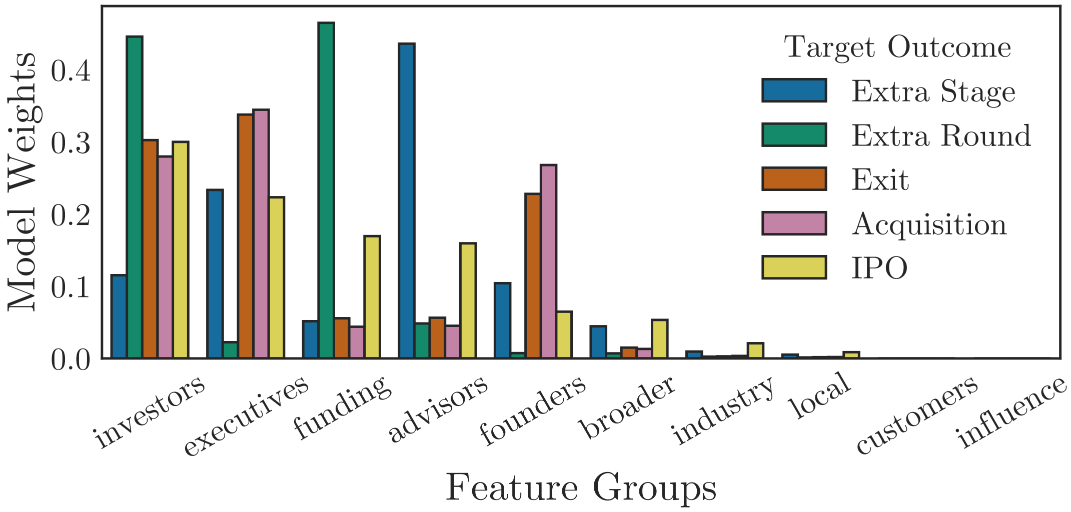


Figure 4.13: Feature weights by target outcome.

CHAPTER 5

Discussion

5.1 Design

5.1.1 Criteria Selection

5.1.2 Feature Selection

5.1.3 Data Sources

5.1.4 Classification Algorithms

5.2 Theory

5.2.1 Efficiency

5.2.2 Robustness

5.2.3 Forecast Window

5.2.4 Developmental Stage

5.2.5 Target Outcome

5.2.5.1 Case Studies

We present four case studies to highlight the nuances of our system's performance, as shown in Table 5.1.

1. ChaCha is an Indiana-based mobile Q&A service, launched in 2005. ChaCha

	Company			
Feature	ChaCha	Doctor.com	Fab	Mixpanel
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
Feature Stage	Series D+	Pre-Seed	Series C	Series A
Outcome Stage	Series D+	Series A	Acquired	Series B
Predicted Outcome	✓	✗	✓	✓
Actual Outcome	✗	✓	✓	✓
Correct Prediction	✗	✗	✓	✓

Table 5.1: Company profiles and predictions.

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 ac-

quired 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 Venture Capital (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. Despite some stumbles in 2016, where MixPanel had to cut 20 staff (primarily in sales), it still appears to have a generally positive outlook. This was a good prediction by our system.

CHAPTER 6

Conclusions

Our project's aim was to produce a Venture Capital (VC) investment screening system that is efficient, robust and powerful. Our system uses online data sources and machine learning to identify startups that are likely to receive additional funding or exit in a given forecast window. While this is a challenging problem even for experienced VC investors, we achieved classification results that have practical application for VC firms.

6.1 Evaluation of Criteria

6.1.1 Efficiency

Our system intends to replace manual investment screening (referral, Google search, industry papers, and manual search of startup databases). Our automated system should be more efficient than these methods because it requires less human input. We also evaluated our system's efficiency based on dataset size and time profile. Our system is relatively robust to dataset size, because it optimises the classification pipeline based on the datasets available. In some cases, our system could use smaller training sets without significant reduction in predictive power. However, our system's ability to use smaller training sets was related to the length of the desired forecast window, and the breadth of the target outcome. An indicative implementation of our system takes 46 hours to run, which is reasonable for a process that is not time-sensitive in this industry. The majority of this time is due to the pipeline optimisation process. However, when placed into production, this could be run less frequently with minimal reduction in performance.

6.1.2 Robustness

Our system must be robust robust in its performance with respect to time so investors can rely on its predictions. The Venture Capital (VC) industry is concerned that predictive models trained on historical data will not accurately predict future trends and activity. This has been identified as a key barrier to the adoption of automated systems by the VC industry [46]. We evaluated our system across a number of historical datasets, forecast windows, and even multiple evaluation metrics. We found variance across all evaluation metrics to be very low, with slightly more variance over shorter forecast windows. When we explored the feature weights for each model developed on different historical datasets, we found only slight variance. This suggests that our system produces highly robust models, suitable for forward-looking investment screening.

6.1.3 Predictive Power

Our system must be accurate at identifying a variety of high-potential investment candidates. We evaluated the systems' predictive power 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). Forecast window has an impact on our system's performance. Our system produced F1 Scores of 0.36, 0.48 and 0.55 for forecast windows of 2, 3 and 4 years. In the future, it would be interesting to also explore forecast windows that are closer in duration to a typical VC investment horizon (5-8 years). Our system's performance has a positive relationship with developmental stage, producing F1 Scores ranging from 0.33 for Pre-Seed companies through to 0.62 for Series C companies. When we fit models separately to each stage, we see a performance improvement (+0.03), and particularly for Series D+ companies (0.53 to 0.66). Finally, our system varies in its performance at predicting target outcomes, producing F1 Scores ranging from 0.51 for predicting additional funding through to 0.24 for predicting an IPO.

6.2 Future Research

6.2.1 Network Analysis

6.2.2 Temporal Analysis

6.2.3 Semantic Text Analysis

6.2.4 Systems Integration

We developed a connector to CrunchBases API that provided real-time and comprehensive access to their database. Although we abandoned this approach because of the time constraints of this research project, it deserves further investigation. In terms of producing a system that is practical for Venture Capital (VC) firms to use day-to-day, continuous data integration is an important feature. The dataset could be continuously updated and incrementally fed into the classification pipeline to a) ensure that the dataset is up-to-date and b) provide more granular temporal analysis of trends in the dataset.

6.3 Summary

We produced a Venture Capital (VC) investment screening system that is efficient, robust and predictive. Existing approaches in the literature to predict startup performance have three common limitations: small sample size, a focus on very early-stage investment, and incomplete use of features. In addition, there is little evidence that previous research has translated into systems that assist investors directly. This project addressed these issues and lays the groundwork for an industry-ready VC investment screening tool.

APPENDIX A

Feature Selection

We develop a conceptual framework relating startup potential and investor confidence to startup investment. We will operationalise this conceptual framework into features that can be incorporated into our machine learning model. To do this, we review features that have been tested in previous studies related to startup investment or performance. In the following sections, we describe each of these features and outline conceptual and empirical evidence that justify their inclusion in our conceptual framework. Figure A.1 depicts how these features can be incoprorated into our conceptual framework.

A.1 Venture Quality

A.1.1 Human Capital

Human capital is critical to early-stage startups that have limited resources and are changing constantly. Startups are composed of founders, non-executive directors (NED) that may be investors or advisers, and staff. Each of these parties makes a contribution to the human capital of the startup. The human capital of these parties can generally be categorised three ways: education, prior experience, and synergies as a team.

Founders' Capabilities Founders play multiple roles in early-stage startups, driving many aspects of the business growth and development. Accordingly, the human capital of founders has been shown to affect startup investment success. In particular, education of founders is a key signal. The number of degrees attained by founders is predictive of success [6, 20], as is whether a founder has obtained an MBA [6]. In addition, past entrepreneurial experience seems to be a predictive factor [20] though there is some evidence to dispute this [41]. Finally, the number of founders seems to be correlated

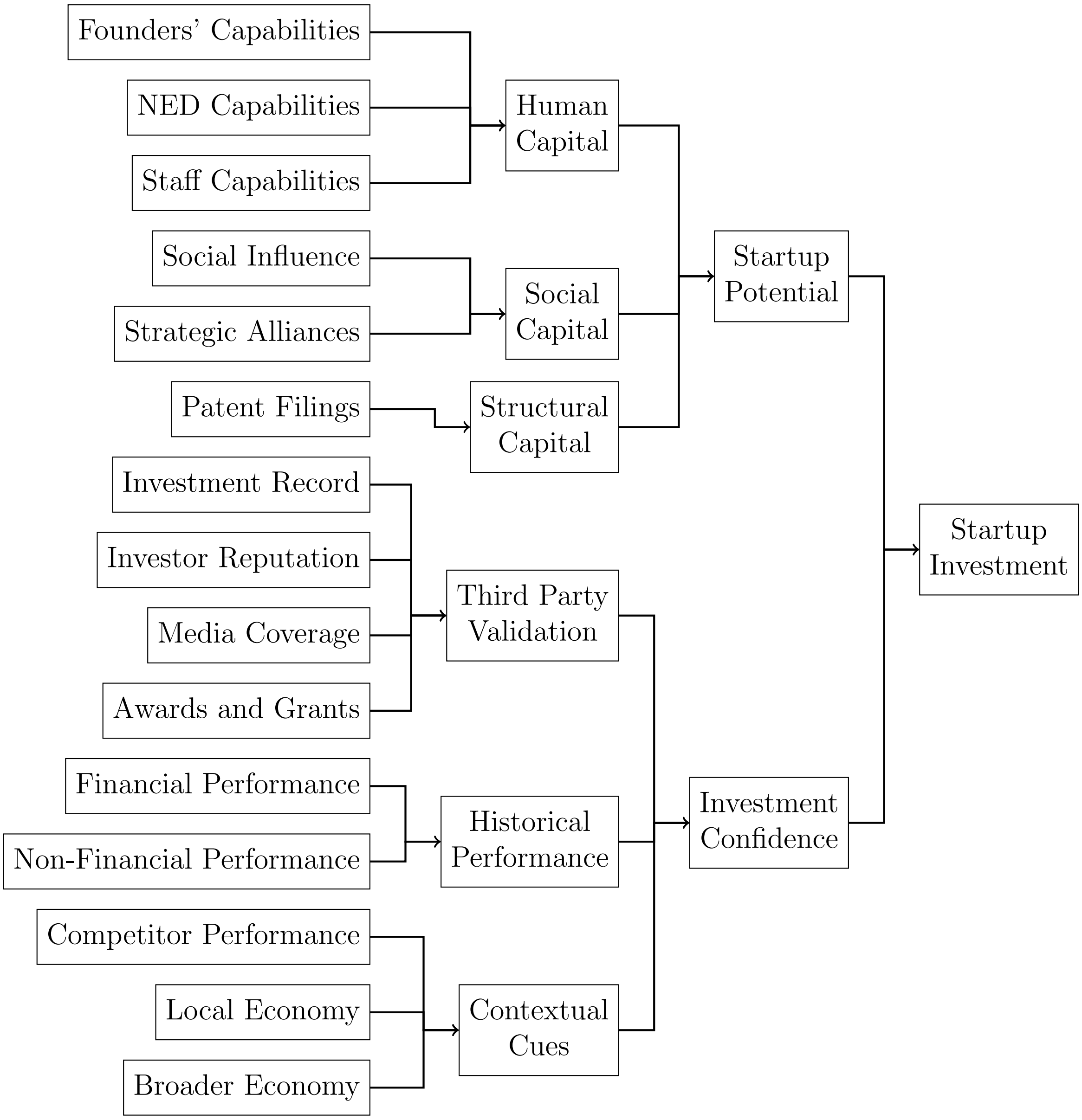


Figure A.1: Proposed conceptual framework for startup investment. This extended version of the framework includes features identified by empirical studies of startup investment. We adapt the framework proposed by Ahlers et al. [1], originally based on work by Baum and Silverman [5].

to startup success [6], though the underlying relationship may be more nuanced, and could be related to the distribution of team skillset.

Non-Executive Directors' Capabilities The boards of startups are smaller and have a higher concentration of ownership than those of well-established companies [26]. Startups lack corporate skills such as finance, human resources, information technology and legal expertise. Especially if founders are relatively inexperienced, they may look to the board to provide these skills. As a result, there is more overlap between governance and operational roles and directors may have greater influence on company performance through greater involvement in decision making [26]. Startups with more experienced directors are more successful at raising funds [5].

Staff Capabilities Founders play a key role in the very early stages of a startup and also in setting the culture for the organisation, but as the organisation grows more importance is given to the influence of employees. Measures like the number of current employees are broad representations of the startup's human capital and are correlated with subsequent startup investment [6, 3, 13]. Detailed analyses of staff human capital are not present in the literature but may be possible using data collected from sources like AngelList and LinkedIn.

A.1.2 Social Capital

Entrepeneurship revolves around opportunity discovery and realisation [43]. Opportunity discovery is only possible through the medium of social networks, so social capital is important. Social networks exist in many forms and contribute in different ways to social capital. These networks can be categorised in terms of the strength of their relationships: weak ties (e.g. social media) and strong ties (e.g. strategic alliances).

Social Influence Startups use social media to communicate with other parties including their customers, potential customers, the media, potential employees, and potential investors. Social media activity can be proxy for a startup's social influence. Startups use different social media platforms for different purposes. Presence and engagement (e.g. number of followers, number of likes, number of posts) on Facebook and Twitter are predictive of startup investment success [12, 6]. These platforms are likely to capture customer or potential customer interactions, which is an indicator of market adoption. In addition, the number of followers on AngelList predicts startup investment success [3], probably because it captures potential employees and investors' interest.

Strategic Alliances Strategic alliances with other companies or institutions have the potential to alter the opportunities that startups can access. Biotechnology startups that have links to industry partners are able to IPO more quickly and at higher market valuations [48]. Startups with more downstream (e.g. manufacturing), but not upstream (e.g. research and development), alliances obtain significantly more venture capital financing than startups with fewer such alliances [5].

A.1.3 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. Structural capital also includes processes and systems but these are less fully-formed in startups than more stable companies.

Patent Filings Many startups develop innovative technologies to help them capture a new market or better capture an existing market. Entrepreneurs protect their ideas through patent filings. Patents are an indicator of the technological capability of the startup. Patents and patent filings affect the survival and investment success of biotechnolgoy startups [5, 24]. However, there may not be as strong a relationship for non-biotechnology startups (e.g. software) [20, 1] This might be because factors like speed-to-market dominate the protective properties of patent filings in the quicker-moving high-technology sector.

A.2 Investment Confidence

A.2.1 Third Party Validation

By their nature, startups are optimsitic about the effectiveness of new technologies and business models. Founders are also highly invested in their startups and therefore it is reasonable for investors to doubt their claims. Third party validation from credible sources like other investors, the media, and the government, may be factored into investors' decision-making process [25, 23].

Investment Record Intuitively, a track record of demand for investment is likely to be a strong signal of future likelihood of future investment. Average funding per round, number of investors per round, number of previous financing rounds and total prior funding raised all predict future likelihood of investment [1, 6, 14, 24, 13].

Investor Reputation Funding from reputable investors sends a clear signal to potential investors that a startup is likely to be of high quality. Investors may believe they require less due diligence because it has been performed by another investor. Startups that receive their initial funding round from a prominent investor are more likely to survive and receive higher valuations in initial public offerings [23]. Followers on AngelList and previous co-investors predict the likelihood of an investor's portfolio startups raising additional rounds successfully [3, 50].

Media Coverage Media coverage provides legitimacy and credibility to startups. Media attention for startups affects the perceived valuation of well-informed experts like venture capitalists [35]. This also translates to increased investment success [6]. There are a few possible explanations for this. First, media coverage signals public interest which might positively influence other stakeholders like customers, employees, etc. Second, new information become widely available which reduces perceived information assymetry.

Awards and Grants Governments and other startup ecosystem supporters often run competitions, grant processes and awards to recognise and celebrate startups. Not only do awards and grants raise the profile of startups but they also indicate third-party validation. Interestingly, there is some evidence that government grants are positively associated with startup investment but awards may have a negative effect [1]. Perhaps this suggests that higher quality startups focus on more critical activities (e.g. raising funds, filing for patents).

A.2.2 Historical Performance

Startup performance is challenging to measure because there are no standardised reporting formats and the availability of data varies wildly. Capturing the multi-dimensionality of startup performance requires the use of multiple measures [51], however, most studies are only able to utilise simplistic performance metrics like survival time [37, 45, 21].

Financial Performance Despite being intuitive, there is little evidence of a relationship between startup financial performance and future investment success. This is because it tends to be difficult to access valid, accurate and complete financial performance measures (e.g. profit, revenue). This information is considered by startups as private and confidential and unlike public companies, private companies are not required to make financial disclosures. Proprietary databases can provide some data on private companies but commercial licenses are expensive and have poor coverage of early-stage companies [4].

Non-Financial Performance With a paucity of financial information available, researchers have looked for other measures of startup performance. Survival time is the most commonly studied startup performance metric despite the coarseness of the measure [45, 3, 20]. There are a few possible explanations for this. One explanation is that startups have such a high failure rate and long time to profitability that many won't ever report any other meaningful performance metrics [39].

A.2.3 Contextual Cues

Startups do not exist in isolation but are rather a product of their context. Investors must consider the performance of a startup's competitors, their local economy and the broader economy when evaluating the reasonableness of signals of startup potential.

Competitor Performance Startups are involved in almost every industry. However, startups across industries have very different requirements, trajectories and measure their performance in different ways. Comparing startups across industries does not necessarily provide a clear view as to whether the potential of a firm is remarkable, likely errant, or within normal ranges. Acccordingly, industry classification has been found to be a key determinant of startup investment [41, 14, 20].

Local Economy Headquarters location is a key indicator of startup investment success [6, 14, 20]. A clear example of this effect is Silicon Valley, a location known for producing an outsized number of successful startups. Silicon Valley provides a focal point for engineering talent, previously successful entrepreneurs, and venture capital firms. Therefore, we might expect different signs of startup potential for Silicon Valley startups compared to those in locations where development and traction are more difficult to attain.

Broader Economy Although startups are less affected by broader economic trends than larger, well-established companies economic challenges have a knock-on effect for startup investment. The Global Financial Crisis led to a 20% decrease in the average amount of funds raised by startups per funding round, disproportionately affecting later-stage funding rounds. Therefore, when comparing startups of different ages, these sort of shocks have key implications for assessing what is a normal trajectory. This may explain why the year a startup is founded can influence startup investment [14, 24].

APPENDIX B

Database Schema

APPENDIX C

Pipeline Hyperparameters

APPENDIX D

Classification Reports

Figure D.1 shows the classification report for each model. The F1-Scores (Positive Class) are identical across the models, at 0.32.

Slice Date		N	(%)	Accuracy	Precision	Recall	F1-Score
2012	Outcome=0	106,583	(90.2%)	-	0.95	0.77	0.85
	Outcome=1	11,538	(9.8%)	-	0.22	0.59	0.32
	Avg/Total	118,121	-	0.76	0.88	0.76	0.80
2013	Outcome=0	106,489	(90.2%)	-	0.95	0.75	0.84
	Outcome=1	11,535	(9.8%)	-	0.21	0.62	0.32
	Avg/Total	118,024	-	0.74	0.88	0.74	0.79
2014	Outcome=0	106,583	(90.2%)	-	0.95	0.76	0.84
	Outcome=1	11,538	(9.8%)	-	0.21	0.61	0.32
	Avg/Total	118,121	-	0.74	0.88	0.74	0.79

Table D.1

Forecast Window		N	(%)	Accuracy	Precision	Recall	F1-Score
2 Years	Outcome=0	319,655	(90.2%)	-	0.95	0.76	0.84
	Outcome=1	34,611	(9.8%)	-	0.22	0.61	0.32
	Avg/Total	354,266	-	0.75	0.88	0.75	0.79
3 Years	Outcome=0	216,816	(85.6%)	-	0.93	0.77	0.84
	Outcome=1	36,602	(14.4%)	-	0.32	0.66	0.43
	Avg/Total	253,418	-	0.75	0.84	0.75	0.78
4 Years	Outcome=0	137,217	(81.9%)	-	0.91	0.81	0.86
	Outcome=1	30,255	(19.1%)	-	0.43	0.65	0.52
	Avg/Total	167,472	-	0.78	0.82	0.78	0.80

Table D.2

Stage		N	(%)	Accuracy	Precision	Recall	F1-Score
Pre-Seed	Outcome=0	491,154	(90.4%)	-	0.94	0.88	0.91
	Outcome=1	51,862	(9.6%)	-	0.27	0.42	0.33
	Avg/Total	543,016	-	0.84	0.87	0.84	0.85
Seed	Outcome=0	136,586	(85.7%)	-	0.93	0.60	0.73
	Outcome=1	22,872	(14.3%)	-	0.24	0.74	0.36
	Avg/Total	159,458	-	0.62	0.83	0.62	0.68
Series A	Outcome=0	23,861	(67.6%)	-	0.87	0.18	0.30
	Outcome=1	11,427	(32.4%)	-	0.36	0.94	0.52
	Avg/Total	35,288	-	0.43	0.71	0.43	0.37
Series B	Outcome=0	12,116	(59.3%)	-	0.84	0.04	0.07
	Outcome=1	8,314	(40.7%)	-	0.41	0.99	0.58
	Avg/Total	20,430	-	0.43	0.66	0.43	0.28
Series C	Outcome=0	5,555	(55.2%)	-	0.82	0.03	0.07
	Outcome=1	4,516	(44.8%)	-	0.45	0.99	0.62
	Avg/Total	10,071	-	0.46	0.66	0.46	0.32
Series D+	Outcome=0	4,416	(64.1%)	-	0.89	0.02	0.03
	Outcome=1	2,477	(35.9%)	-	0.36	1.00	0.53
	Avg/Total	6,893	-	0.37	0.70	0.37	0.21

Table D.3

Target Outcome		N	(%)	Accuracy	Precision	Recall	F1-Score
Acquisition	Outcome=0	744,920	(96.1)	-	0.98	0.96	0.97
	Outcome=1	30,236	(3.9)	-	0.37	0.61	0.46
	Avg/Total	775,156	-	0.94	0.96	0.94	0.95
Exit	Outcome=0	738,564	(95.3)	-	0.98	0.96	0.97
	Outcome=1	36,592	(4.7)	-	0.42	0.60	0.49
	Avg/Total	775,156	-	0.94	0.95	0.94	0.95
Extra Round	Outcome=0	694,804	(89.6)	-	1.00	0.81	0.89
	Outcome=1	80,352	(10.4)	-	0.37	0.99	0.54
	Avg/Total	775,156	-	0.83	0.93	0.83	0.86
Extra Stage	Outcome=0	673,688	(86.9)	-	0.93	0.84	0.89
	Outcome=1	101,468	(13.1)	-	0.36	0.60	0.45
	Avg/Total	775,156	-	0.81	0.86	0.81	0.83
IPO	Outcome=0	768,207	(99.1)	-	0.99	0.99	0.99
	Outcome=1	6,949	(0.9)	-	0.26	0.39	0.31
	Avg/Total	775,156	-	0.98	0.99	0.98	0.99

Table D.4

Bibliography

[1] Ahlers, G. K., Cumming, D., Gunther, C., and Schweizer, D. “Signaling in equity crowdfunding”. In: *Entrepreneurship Theory and Practice* 39.4 (2015), pp. 955–980.

[2] Ahn, H. and Kim, K.-j. “Using genetic algorithms to optimize nearest neighbors for data mining”. In: *Annals of Operations Research* 163.1 (2008), pp. 5–18.

[3] 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.

[4] Artemchik, T. “PrivCo”. In: *Journal of Business & Finance Librarianship* 20.3 (2015), pp. 224–229.

[5] 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”. In: *Journal of Business Venturing* 19.3 (2004), pp. 411–436.

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

[7] 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. Berlin: Springer, 2011, pp. 399–410.

[8] Breiman, L. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[9] Bulmer, M. G. *Principles of statistics*. Courier Corporation, 1979.

[10] 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.

[11] 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.

[12] Cheng, M., Sriramulu, A., Muralidhar, S., Loo, B. T., Huang, L., and Loh, P.-L. “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] Conti, A., Thursby, M., and Rothaermel, F. T. “Show Me the Right Stuff: Signals for High-Tech Startups”. In: *Journal of Economics & Management Strategy* 22.2 (2013), pp. 341–364.

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

[15] 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.

[16] Dixon, M. and Chong, J. “A Bayesian approach to ranking private companies based on predictive indicators”. In: *AI Communications* 27.2 (2014), pp. 173–188.

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

[18] Fried, V. H. and Hisrich, R. D. “Toward a model of venture capital investment decision making”. In: *Financial management* (1994), pp. 28–37.

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

[20] Gimmon, E. and Levie, J. “Founder’s human capital, external investment, and the survival of new high-technology ventures”. In: *Research Policy* 39.9 (2010), pp. 1214–1226.

[21] Gloor, P. A., Dorsaz, P., Fuehres, H., and Vogel, M. “Choosing the right friends–predicting success of startup entrepreneurs and innovators through their online social network structure”. In: *International Journal of Organisational Design and Engineering* 3.1 (2013), pp. 67–85.

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

[23] Hochberg, Y. V., Ljungqvist, A., and Lu, Y. “Whom you know matters: Venture capital networks and investment performance”. In: *The Journal of Finance* 62.1 (2007), pp. 251–301.

[24] Hoenen, S., Kolympiris, C., Schoenmakers, W., and Kalaitzandonakes, N. “The diminishing signaling value of patents between early rounds of venture capital financing”. In: *Research Policy* 43.6 (2014), pp. 956–989.

[25] 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.

[26] Ingley, C. B. and McCaffrey, K. “Effective governance for start-up companies: regarding the board as a strategic resource”. In: *International Journal of Business Governance and Ethics* 3.3 (2007), pp. 308–329.

[27] Kotsiantis, S. “Supervised Machine Learning: A Review of Classification Techniques”. In: *Informatica* 31.3 (2007).

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

[29] Liang, Y. E. and Yuan, S.-T. D. “Predicting investor funding behavior using crunchbase social network features”. In: *Internet Research* 26.1 (2016), pp. 74–100.

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

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

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

[33] 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.

[34] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. “Scikit-learn: Machine learning in Python”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.

[35] Petkova, A. P., Rindova, V. P., and Gupta, A. K. “No news is bad news: Sensegiving activities, media attention, and venture capital funding of new technology organizations”. In: *Organization Science* 24.3 (2013), pp. 865–888.

[36] Puschmann, C. and Burgess, J. “The politics of Twitter data”. In: (2013).

[37] Raz, O. and Gloor, P. A. “Size really matters-new insights for start-ups’ survival”. In: *Management Science* 53.2 (2007), pp. 169–177.

[38] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning representations by back-propagating errors". In: *Cognitive Modeling* 5.3 (1988), p. 1.

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

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

[41] 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.

[42] Shane, S. and Cable, D. "Network ties, reputation, and the financing of new ventures". In: *Management Science* 48.3 (2002), pp. 364–381.

[43] Shane, S. and Venkataraman, S. "The promise of entrepreneurship as a field of research". In: *Academy of Management Review* 25.1 (2000), pp. 217–226.

[44] Shin, K.-S., Lee, T. S., and Kim, H.-j. "An application of support vector machines in bankruptcy prediction model". In: *Expert Systems with Applications* 28.1 (2005), pp. 127–135.

[45] Song, Y. and Vinig, T. "Entrepreneur online social networks–structure, diversity and impact on start-up survival". In: *International Journal of Organisational Design and Engineering* 2.2 (2012), pp. 189–203.

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

[47] 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." In: *Psychological Methods* 14.4 (2009), p. 323.

[48] Stuart, T. E., Hoang, H., and Hybels, R. C. "Interorganizational endorsements and the performance of entrepreneurial ventures". In: *Administrative Science Quarterly* 44.2 (1999), pp. 315–349.

[49] 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.

[50] Werth, J. C. and Boeert, P. "Co-investment networks of business angels and the performance of their start-up investments". In: *International Journal of Entrepreneurial Venturing* 5.3 (2013), pp. 240–256.

[51] Wiklund, J. and Shepherd, D. "Entrepreneurial orientation and small business performance: a configurational approach". In: *Journal of Business Venturing* 20.1 (2005), pp. 71–91.

[52] 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.

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