Predicting Start-up Success with Machine Learning: Analysing the Effect of Network-Based Features

James Ellis

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2021

Abstract

In this dissertation, I explore two network features from the literature: an investor syndicate network, and a network of professional relations among companies. Using several different centrality measures, I explore the partial contribution of each network-based feature in a binary classification task - predicting start-up success. In particular, due to recent excellent results in the literature, I chose XGBOOST as the classifier for this project. Using the Crunchbase database, I construct a dataset of 81 independent variables, containing 161,759 early-stage start-ups, with funding prior to Series A, to build the model from. Models incorporating the network-based features were constructed and evaluated with a variety of metrics to compare and contrast performance. By deploying SHAP analysis, I show the network of professional relations to be a more important feature than the syndicate network. Centrality values in the former were positively correlated with feature contribution towards the prediction of a successful outcome; the latter also showed similar correlations up to a peak, leading into a plateau. This raised questions about the benefit of high status, well-connected investors in funding rounds prior to Series A. These findings launch a discussion, along with suggestions for future research avenues for the benefit of the interested reader.

Acknowledgements

I would first like to thank Dr. Filippo Menolascina for his continual support in what has proved to be a very interesting field. I also extend my appreciation to Dr. Lucia Bandiera for advice on data-processing. Finally, I would also like to thank Dr. Moreno Bonaventura and Dr. Javier Arroyo for valuable communications which clarified certain nuances in their methodologies.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*James Ellis*)

Table of Contents

Chapter 1

Introduction

Investing in start-ups is a notoriously high-risk activity, particularly in the early-stages of their life cycle [1]. Even though there is an overwhelming trend for a start-up to fail, there is still potential for investors to make a profit due to an asymmetry between the potential losses and gains. Should a start-up fail, the most an investor can lose is the initial investment; however, should they invest in the next Google or Facebook, their initial investment will multiply many times as it profits off the exponential growth that characterises the successful minority.

Venture capitalists (VCs) are a form of private equity investors that focus on funding and nurturing promising start-ups. Their goal is to provide capital and support until the start-up is deemed successful, at which point, the VC will profit. In order to source the deals in the first place, VCs normally use their professional networks [2], or by start-ups directly contacting the VC. These deals then require a screening process, where the VC determines the suitability of the deal. Many, if not all, VCs rely on their own intuition, domain-specific knowledge, and ”gut feel”. Not only is the screening process resource intensive, but it also highly subjective and affected by several forms of bias [3, 4, 5].

Recently, there has been an increasing amount of competition amongst VCs. Whilst the rate of start-up creation has remained at a near constant rate [6], the number of VC funds (and the size of the funds) has been increasing [7]. This has increased time pressure on VCs to identify the best deals, increasing the risk of poor investment decisions. Moreover, there has been a recent trend of VCs turning to quantitative methods for sourcing [8], such as web scraping, increasing the overall deal flow. However, screening practices have fallen behind the pace and are more or less the same since the inception of the VC sector over 80 years ago. The current situation for VCs is clear:

increasingly high quantities of deal flow are being processed by increasingly time pressured and biased screening methods. This is a situation which risks the profitability of the VC. In this project, I turn to machine learning to contribute towards solving this bottle-necked process.

In the financial industry as a whole, artificial intelligence (AI), and in particular machine learning (ML), has widely been utilised due to the vast amounts of quantitative data that largely characterise the industry. For example, ML has been applied to detect credit card fraud [9] and to build quantitative trading models [10]. Successful applications of ML, like the aforementioned examples, have made the financial industry the second largest sector by expenditure on AI services [11]. However, venture capital is lagging behind the rest of the industry in ML uptake, with estimates suggesting only 5% of the industry is using some form of AI [12].

From an academic viewpoint, this hesitancy in ML uptake is quite surprising. There is a healthy amount of literature that directly targets ML in VC. There are also several start-up databases [13], rich with details on their founders, investors, and other valuable information. On the other hand, there are several reasons from a VCs viewpoint that explains the hesitancy. Firstly, many of the key criteria a start-up is assessed on are intangible, such as the quality of the entrepreneur(s) or the quality and timing of the idea. Secondly, anecdotal evidence [14] amongst VCs has shown a fear that algorithms will miss out on the next Google or Facebook, consequently affecting the profitability of the VC.

In an attempt to alleviate these concerns, I follow a hybrid approach suggested in prior research [14]. In particular, ML is used to target the early stages of the screening process to reduce the amount of deal flow. VCs can then spend more time evaluating a reduced number of potential deals, putting themselves in the best position possible to secure deals in a competitive market. In this way, ML acts to remove a massive time burden from the manual screening process, but VCs don't give up full control over the decision process, since any further screening and decision making is made by the VC. There is very low tolerance for an algorithm which removes companies that would eventually be successful. Therefore, the addition of any features which could create better screening algorithms would be of benefit to both the academic and VC community.

The aim of this project is to investigate the partial contribution of two network-based features from the literature: an investor's centrality in a syndicate network, and a start-up's centrality in a network of professional relations among companies. Both

have shown to be statistically significant features correlated with economic success, hence they are good candidates for an analysis as part of a wider feature set in ML classification. Consequently, this project explores two key questions: do the network-based features improve the performance of a classifier with a baseline set of features? If so, how important are they relative to the other features in the algorithm?

To answer these questions, I use the Crunchbase database to develop a dataset of over 160,000 start-ups, with 81 features ranging from company location to information about the founders' education. Using XGBOOST, the current state-of-the-art in gradient boosted tree classifiers, I measure several model metrics, investigating any changes when incorporating the network features. Furthermore, using SHAP analysis, I explore the importance of each network-based feature and provide an analysis of how the importance varies with feature value. Experimental results show both features to be within the top 15 most important features, with the professional relations network centrality proving to be more significant to the classification. With further analysis, I find that the centrality of a start-up in the professional relations network is positively correlated with a successful outcome. A similar trend is observed with the syndicate network centrality up to a peak value, with a counter-intuitive plateau then observed. Although an in-depth analysis of the trend was out of the scope for this project, I launch a discussion with directions for future work which could investigate this finding.

The remainder of this paper is organised as follows: Chapter 2 covers the relevant background material and literature whilst also developing the hypotheses of the project. Chapter 3 provides an in depth and reproducible summary of the methodology, whilst Chapter 4 presents the results and launches a discussion about the findings. Finally, we finish with Chapter 5 which details the limitations of the project, suggestions for future work, and concluding remarks.

Chapter 2

Background and Hypothesis Development

As mentioned in the introduction, the goal of this work to determine the partial contribution of two network-based features as part of a wider feature set in ML classification. Before providing any hypotheses, I first provide a brief background into the different funding rounds a venture might incur during their life cycle. Further to this, I provide an overview of the stages of a venture capitalist prior to an investment. The more inclined reader may want to skip directly to Sec 2.3, where we explore relevant literature and layout hypotheses which I ultimately try to prove in the later chapters of this dissertation.

2.1 Funding Life-cycle of a Start-up

From an entrepreneur's idea, to the potential of a very successful company, the start-up is likely to go through many different stages of funding to facilitate its growth. Some of the different stages you might see, in chronological order, include[1]:

- **Pre-seed** The first funding required to convert an idea into a start-up. It is generally provided by the founders, friends, and family. Typically, there is no equity exchange, or in other words, the friends and family don't receive partial ownership in the venture.
- **Seed** Generally, this is the first funding where there is an exchange of capital

[1] By no means is this a formal or exhaustive list. Informal definitions have been presented for the benefit of the uniformed reader.

from investors for partial ownership in the start-up. Angel investors and incubators/accelerators are the typical investors at this stage.

- **Series A** At this point, the company has a developed product with revenue flow. This funding helps to scale the product. This is typically the first stage where early-stage VCs will invest.
- **Series B** This funding, which is mostly provided by VCs, can be used to strengthen the work force or increase product exposure to new customers.
- **Series C** At this stage, the start-up is well established, has a good history of stable revenue and growth, with a desire to expand to even wider audiences, or even operate at the global level. The typical investor will be a late-stage VC.

Most start-ups will end up failing at varying points in the funding life-cycle; however, as a start-up progresses through the funding stages, the risk associated with that start-up failing decreases [1]. A start-up could fail for a number of reasons such as lack of funds, or inability to secure further investment because, although the idea may be good, the plan to monetize is not viable. The successful minority do not necessarily have to follow this funding route. For example, some start-ups may completely skip certain funding rounds, and others may only need up to Series A before being considered a success. On the other hand, funding has the potential to go further depending on the economic needs and end goals of the business. The size of the investment also tends to get larger at later stages of investment. For example, from the Crunchbase database, the average seed investment is $1.1M, whereas Series C is over thirty times larger at $32.1M.

2.2 Stages of Venture Capital

A useful analogy for the VC investment process is to think of potential deals flowing into a funnel: the top of the funnel will receive lots of potential deals (deal flow), which will eventually be whittled down to the few the VC ultimately invests in. In a more theoretical framework, Fried and Hisrich [15] describe the decision making of a venture capitalist in a six stage model: sourcing, two stages of screening, two stages of evaluation, and finally closing the deal. The VC will potentially profit off any closed deal should the start-up eventually be acquired by another company (acquisition), or

be put on a public stock market (IPO). Either of these events provide a way for the VC to liquidate their shares in the start-up.

2.3 Machine Learning in Venture Capital

There has been a recent trend for VCs to use quantitative methods in the sourcing stage [8]. The main edge quantitative sourcing provides is to find high potential deals quickly, putting the VC in a better position to secure the deal. Quantitative sourcing involves web scraping websites such as Companies House for new company registrations, Twitter for trending companies, or by periodic monitoring of seed investor websites for recent investments. Whilst sourcing methods appear to be modernising, screening methods are more or less the same since the inception of VC as a sector just after World War II. Not only are current screening methods resource intensive, but they are also subjective and influenced by biases such as overconfidence bias [3], similarity bias [4], and availability bias [5]. With increasing amounts of competition, VCs find themselves in a position were they need to identifying potential deals quickly, whilst also having to deal with increased deal flow should they have quantitative sourcing methods in place. This is pushing current screening methods to their limit, and runs the risk of poor decision-making. Therefore, the screening stage is an ideal candidate for ML to reduce the amount of manual screening required. This would allow VCs to spend more time manual screening potentially ”good” deals, thus allowing for better decision-making to be made in this ML-human hybrid decision process. Consequently, the work in this project targets early-stage VCs, particularly those who have implemented quantitative sourcing.

2.4 The Transfer of Knowledge

Bonaventura et al. [16] showed that there is a statistically significant link between the professional relations a start-up has formed and the likelihood of a positive economic outcome. The idea is that, if an employee from a large and successful company, such as Google, moves to a start-up, the start-up will gain some of Google’s ”know-how”. For example, a skilled employee could bring with them knowledge of cutting-edge technology, or seasoned advisors can provide valuable insight on business strategy. By constructing a network of companies, linked by the transfer of employees amongst the former, which Bonaventura et al. call the *WWS Network*, the most important start-ups

can be identified by their centrality in the network (discussed further in Sec 3.2). The more central a start-up is in this network, the higher the likelihood success. Informed by the aforementioned results, I hypothesise the following:

Hypothesis 1a: *Start-up centrality in the WWS network will be positively correlated with the former's contribution towards positive* [2] *model predictions.*

Hypothesis 1b: *The addition of start-up centrality in the WWS network to a baseline model will provide an improvement of the measured metrics.*

2.5 Investor Knowledge and Reputation

Start-up investors provide much more than just capital. Investors may also partake in activities, such as addressing weaknesses in the entrepreneurial team [17], or transforming the start-up into a professional company [18]. Hochberg et al. [19] showed that start-ups associated with better networked investors are more likely to experience a successful outcome. In particular, they look at VC networks formed through co-investments with one another. This is known as a syndicate network. A syndicate forms when investors make a joint decision to share the investment of a start-up and has several benefits such as: sharing risk; increasing portfolio diversification [20]; better investment decision-making [21]; and syndicate partners complementing each others knowledge and resources to help nurture the start-up [22]. Hence, the more central an investor is in a syndicate network, the more access to expertise the former has, from which the start-up can benefit. Moreover, highly central investors also represent high social capital that is mutually respected by the other investors [23, 24]. Start-ups that associate with a high centrality investor have their quality certified by the investor's high reputation [25]. VCs are well known to syndicate [21]; however, it is typical to see syndication amongst other types of investors too, such as angel investors [26].

An ML screening algorithm could see value in more central investors in the syndicate network for two reasons. First, the start-up will benefit from the increased access to better knowledge and expertise. Second, highly central investors are more likely to pick better investments to protect from tarnishing their reputation. Hence, I form the following hypotheses:

[2] Positive model prediction refers to the classifier predicting "successful" for a given data input.

Hypothesis 2a: *Investor centrality in the syndicate network will be positively correlated with the former's contribution towards positive model predictions.*

Hypothesis 2b: *The addition of investor centrality in the syndicate network to a baseline model will provide an improvement of the measured metrics.*

Chapter 3

Methodology

In this chapter, I detail the methodology followed in order to obtain the results presented in chapter 4. First, I discuss the dataset, along with the wrangling procedure followed in order to obtain the features used in the classifier. Furthermore, I discuss the ML method used, layout the procedure followed to optimise and train the classifier, and provide an overview of the metrics used. Brief definitions and equations will also provided to complement the methodology.

3.1 Data and Data Wrangling

3.1.1 Crunchbase Database

In this project, I was given access to the Crunchbase (CB) database dated up to 15th December 2020. Of interest to this project, the database contains: information on people, such as jobs or degrees they have achieved; information on investments, such as funding rounds the organisation has gone through or the investment history of the investors; information on acquisitions and IPOs, such as the date they were announced. This information is spread among 17 .csv files, but it is not guaranteed that organisations in the database will have as much information as each other. For example, some organisations may have no information on their founder(s).

Crunchbase offer both the .csv export and access to the API. One distinction between the two is that the API offers access to trust codes. In essence, trust codes give information on how reliable a date is. For any date field in the database (e.g the date a company was founded on), dates will default to the 1st January if only the year is provided. Similarly, dates will default to the 1st day of a month if only the year and month

is provided. Trust codes provide a way to distinguish between genuine and defaulted dates. However, I had to make the assumption that all dates were reliable because I did not have access to the API during this project.

3.1.2 Warm-up and Simulation

Due to the temporal nature of the data, creating a naive train-test split of the entire CB database would not be realistic in a VC setting. Instead, I follow Arroyo et al. [27] who use two distinct time windows:

- **Warm-up Window** ($t_c \leq W < t_s$) This is the period before the VC has made their investment, and can be considered the information they would have access to prior to any investment decision. The company must be created within this window, and not undergone acquisition or an IPO. Arroyo et al. also impose that the start-up must only have received funding prior to Series C, or none at all. I adapt this criteria for funding to be prior to Series A, or not at all. This is better suited to target the screening process of early-stage VCs.
- **Simulation Window** ($t_s \leq W < t_f$) The point at which this window starts, t_s, I assume the VC has made their investment decision. If any of further funding rounds, acquisition, or an IPO occur within the simulation window, the start-up is considered a success. In the case none of these occur within the simulation window, I consider the start-up a failure. This binary success or failure variable is the dependent variable which the classifier attempts to predict.

Not only is this approach more realistic, but it also helps to alleviate survivorship bias in the CB database. The CB database was founded in 2007, therefore any companies that failed prior to this date are much less likely to be present than those that succeeded. By using the most recent data available, this bias can be minimised. Hence, I define a 4 year warm-up window, starting on $t_c = 15^{th}$ December 2013. The simulation is defined over a 3 year period, ending on the last date available in the database, $t_f = 15^{th}$ December 2020.

3.1.3 Baseline Features

I follow Arroyo et al. to create the baseline features which were used throughout the entire project. The main reason for choosing their approach is because their methodology is reproducible and they only use the CB database for their features. Often in

the literature, researchers complement their dataset with features from other databases such as Pitchbook (e.g [14]). However, this was not an option given the very high cost of obtaining such databases [13]. Some features in the CB database have no temporal information; for example, `employee_count` is accurate as of 15^{th} December 2020, and I do not have access to a value accurate at t_s. Features without temporal information have been excluded from the feature set. The features included can be split up into three distinct groups: company information, funding information, and founder information. In total, there are 81 baseline features for 161,759 companies in the warm-up window. All start-ups in the dataset have company information, whilst only 29,364 and 33,741 have funding and founders information respectively.

Company Information

Firstly, any organisations with a `primary_role` of investor or school were removed from the dataset. I then 1-hot encode the `country_code` which represents the location of the company, with any entries with missing entries removed. This led to the addition of more than 200 features. To alleviate the ill-effects of the *curse of dimensionality*, I reduce the number of countries to the top 9 countries with the largest number of billion dollar start-ups [28], with another dummy variable representing all other countries. Organisations also state the category (of which there can be more than one) that best matches the service or product they provide. The 47 unique categories in `category_groups_list` were 1-hot encoded, with any empty entries removed from the dataset. Arroyo et al. further calculate the age of the company in months. However, given the absence of trust codes, I calculate the age of the company in years (`age_years`) instead. Fig. 3.1 shows the number of companies against `founded_on` date. There are large spikes at the start of each year, with secondary maxima observed at the start of each month. This implies that many `founded_on` dates are only accurate to the year, hence a company age variable at a month-level quantisation would be biased. Finally, should the company have social media presence or contact information, I indicate with the presence of a binary variable in the features `has_linkedin`, `has_facebook`, `has_twitter`, `has_phone`, and `has_email`.

Funding Information

As mentioned in the warm-up window definition, I only consider funding rounds prior to Series A. Hence, only pre-seed, seed, or angel are considered for

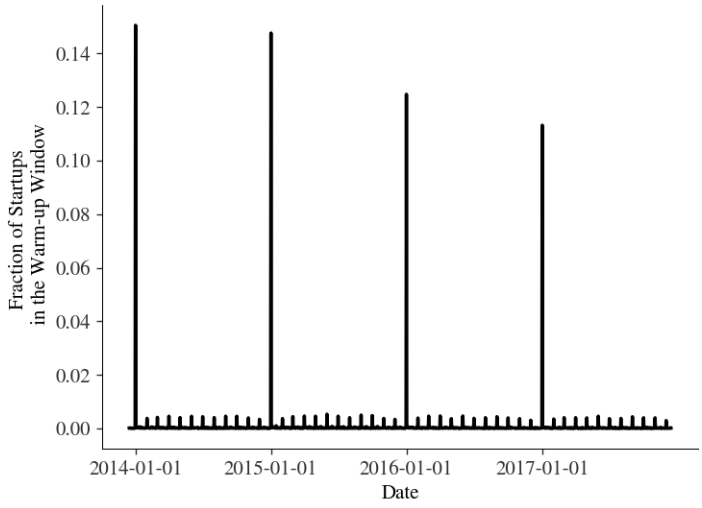

Figure 3.1: A plot of the normalised count against founded on date for start-ups in the warm-up period.

`investment_type`. Generic information about funding in the warmup period is added such as the total number of funding rounds (`round_count`), and the total amount of capital raised in USD (`total_raised_amount_usd`). Further to this, information on the last funding round in the warm-up window, such as amount raised in USD (`last_round_raised_amount_usd`), the type of funding round, which was 1-hot encoded, and the number of months elapsed since the funding round (`last_round_timelapse_months`) were included. I consider a month-level quantisation because funding round dates are typically more accurate due to investor announcements or news articles. Furthermore, I add information on the investors, such as the number of unique investors (`total_investor_count`) along with the number of investors in the last round (`last_round_investor_count`). Following Arroyo et al., I define a ”known” investor as one with their information present in the CB database[1]. With this definition, I further include the number of unique ”known” investors in the warm-up period (`known_investor_count`) and the number of known investors in the last round (`last_round_known_investor_count`).

[1] Arroyo et al. describe these investors as ”renowned”. However, this adjective is somewhat misleading given the context of this project.

Founder Information

In this set of features, I was only interested in the founders of a company, hence removed all people with a `featured_job_title` that didn't contain the string "founder". Using this information, I obtain the number of founders in each company (`founders_count`) along with the a count of male and female founders (`founders_male_count` and `founders_female_count`). Having removed any people with an unspecified nationality, I add country diversity among the founders by counting the number of different nationalities (`founders_dif_country_count`). The founder information was then further complemented with information on their education. I inspected each founder's degree(s), in particular, only considering completed degrees, and removing any empty or unknown `degree_type`'s. Some founders have achieved multiple degrees, but contained in a comma-separated free-form text. Strings were compared to a database of degree abbreviations, along with less strict matching criteria for words such as "bachelor" or "master" (see Appendix A.1 for a full list of the search criteria used). With this procedure, degrees were classified as Bachelors, Masters, or PhD and 1-hot encoded (`has_bachelors`, `has_masters`, `has_phd`).

3.2 Network Features

3.2.1 Graph Theory and Centrality Measures

Formally, any network can be defined as a graph. A graph, $G = (V, E)$, is a mathematical structure that defines pairwise relationships between the constituent elements of the network. The latter are known as nodes or vertices, V, whilst the former are known as edges, E. In an undirected graph, relations between nodes have no direction; if there is connection between node i and j, there is also a connection between j and i. This is not the case in a directed graph where connections are assumed to have a direction. Furthermore, loops - an edge between the same node - is not allowed. Hence, this leads to the formal definition of an edge being defined as $E \subseteq \{\{x, y\} \mid x, y \in V \text{ and } x \neq y\}$. Graphs can be represented by an $n \times n$ matrix, A, called the adjacency matrix. Elements of the adjacency matrix, $A_{ij} \in \{0, 1\}$, act as a binary variable indicating the presence of an edge between node i and j. For undirected graphs, the adjacency matrix is symmetric i.e. $A_{ij} = A_{ji}$.

A special type of graph, called a bipartite graph, $G = (U, V, E)$, is one in which the nodes can be split into two disjoint sets, U and V. Edges can only be formed across the

two sets. This type of graph is particularly useful when there are two different groups of objects. However, to measure the centrality of a node within a specific group, the bipartite graph is first projected onto the same group. For example, if $u_1, u_2 \in U$ is connected to $v_1 \in V$, then the bipartite projection onto U will result in a connection between u_1 and u_2.

In graph theory, centrality measurements are used to evaluate the importance of a node in a network. There are many different measurements of centrality, some of which may give differing results. This project focuses on four main measures, which I define for node $i \in V$ of the graph $G = (V, E)$:

Closeness [29] measures the reciprocal of the distance to all other nodes in the network. The closer the node is to all other nodes, the higher the centrality value of that node,

$$C_c(i) = \frac{n-1}{\sum\limits_{j \in V \setminus \{i\}} d(j,i)} \tag{3.1}$$

where n is the number of nodes in the network, with $n-1$ acting to normalise the metric. The (geodesic) distance, $d(j,i)$, is the number of edges traversed in the shortest path between node j and i. In practice, algorithms slightly amend Eq. 3.1 to account for disconnected nodes [30].

Betweenness [31] measures the extent to which others rely on a specific node to make connections in the network. Or in other words, the extent to which a specific node acts as an intermediary for the connection of other nodes.

$$C_b(i) = \sum_{j,k \in V \setminus \{i\}} \frac{\sigma(j,k \mid i)}{\sigma(j,k)} \tag{3.2}$$

where $\sigma(j,k)$ represents the total number of shortest paths from node j to k, whilst $\sigma(j,k \mid i)$ represents the former but conditioned on passing through node i. Similarly, betweenness is also normalised, but not included for sake of brevity [32].

Degree is a basic measure in graph theory. It is simply the number of nodes i is connected to, otherwise known as the neighbours of i. As a centrality measure, it is normalised by the highest theoretical value of degree in the network, $n-1$ i.e. every node, other than i, connected to i.

$$C_d(i) = \frac{\sum\limits_{j \in V \setminus \{i\}} A_{ij}}{n-1} \tag{3.3}$$

Eigenvector [33] is similar to degree centrality, but takes into account the centrality of neighbouring nodes in a recursive fashion. The centrality value is given by

$$C_{ev}(i) = x_i \tag{3.4}$$

where x_i is the i^{th} element of the eigenvector, $\mathbf{x}$, with eigenvalue, λ, which solves the equation[2]

$$A\mathbf{x} = \lambda\mathbf{x} \tag{3.5}$$

where A is the adjacency matrix of $G = (V, E)$.

3.2.2 Syndicate Network

To create the syndicate network, information on company funding rounds and their investors were sorted. Since I am only interested in syndicates, funding rounds provided by solo investors were removed. I also only consider funding rounds that occurred between the end of the warm-up window, and 5 years prior. This follows Hochberg et al. [19] who found stronger correlations with positive economic outcomes using short time windows. As such, any entries with missing funding round dates were removed. A bipartite graph with investors and funding rounds as the two disjoint sets was constructed, with edges representing an investor's participation in a syndicated funding round. A projection onto the investors was then performed from which all four centrality metrics could be measured.

Unlike the work in this project, Hochberg et al. considers directed centrality measures. They use the former as a way to quantify the amount an investor invites or has been invited to join a syndicate. However, Hochberg et al. uses a different dataset to this work. Although the CB database does contain a binary field indicating the lead investor of a syndicate, over half of the entries are missing. For risk of introducing bias into the metric, paired with the fact that Hochberg et al. found their best performing centrality to be undirected, I decided to exclude these centrality measures.

[2] In practice, power iteration is used to find the eigenvector with the largest eigenvalue. The Perron-Frobenius theorem gaurentees the eigenvector has strictly positive components, given that the adjacency matrix is real, positive, and square.

3.2.3 WWS Network

To construct the WWS network introduced by Bonaventura et al. [16], information on jobs was sorted. Jobs with a starting date that occurred before the company was formed, or without a starting date provided were removed. Only jobs that started prior to t_s were considered. Since the WWS network focuses on the movement of people, those that never moved company were removed. Following the same procedure as the syndicate network, a bipartite graph was formed with people and companies as the disjoint sets, with edges representing a person working or having worked for a company. A projection onto the company nodes was performed, from which all four centrality metrics could be measured.

3.3 Model

3.3.1 XGBOOST

The main aim of this project is to investigate the importance of the two network-based features, not to compare and contrast different classifiers. A common theme amongst recent research is XGBOOST ranking top in most model metrics [27, 14], and hence was chosen as the classifier in this project. XGBOOST [34] is a regularised gradient boosted tree model that performs particularly well with tabular data, and can automatically handle missing values. In short, the algorithm adds together many regression trees to form an ensemble. Each additional tree added to the ensemble is chosen such that it minimises an objective function of the form:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{3.6}$$

where θ are the model parameters. The loss function, $L(\theta)$, measures the difference between predictions and the actual classes of the data. The regularisation function, $\Omega(\theta)$, gets larger with more complex models, hence minimising this term helps to prevent over-fitting. In this project, I use the log-loss function defined as:

$$L(\theta) = \sum -y \log(\hat{y}) - (1-y) \log(1-\hat{y}) \tag{3.7}$$

where y is the actual class label and $\hat{y}$ is the model prediction.

3.3.2 Training

In this project, there are four distinct experiments: the baseline; the baseline and syndicate network; the baseline and Bonaventura's WWS network; and the baseline with both networks. The baseline model uses the features discussed in Sec 3.1.3. The next two experiments have four unique models each, to account for the different centrality measures considered in this project. Using model metrics from these two experiments, the best performing centrality measures were used to create the model incorporating both networks into the feature set.

For each model, the data was separated into a stratified training and test set with a 75:25 split. Stratification ensures that the distribution of classes within the dataset is the same among the splits. Each model used the same random seed for identical training and test sets. The test set was held-out and only used for final evaluation of each model. The training set was used to find the optimal hyper-parameters, and then ultimately train the model. To account for the class imbalance, in which 78.7% of start-ups are unsuccessful, the successful data points were weighted by the ratio of unsuccessful to successful start-ups in the dataset.

For the hyper-parameter search, the training data was further split in a stratified fashion to facilitate a 3-fold cross-validation scoring of each set of hyper-parameters. Cross-validation helps to alleviate any erroneously high validation scores from "lucky" splits of the data. Hyper-parameters were found using the tree-structured parzen estimate algorithm [35][3]. In short, this method leverages Bayesian statistics to make an informed decision on each successive hyper-parameter trial. This is in contrast to popular methods such as random hyper-parameter search, which randomly samples hyper-parameters over the defined parameter space. For each model, the set of hyper-parameters which maximised the cross-validated ROC-AUC score were chosen. This entire process was repeated three times in total, each with a different random seed for the training-test split. The mean of model metrics were taken across the three repeats, with the standard errors, α, given by

$$\alpha = \frac{\sigma_{n-1}}{\sqrt{n}} \tag{3.8}$$

where σ is the (sample) standard deviation of the metric values and n is the number of times the model was repeated using a different split of the data.

[3]The parameter space specified in this project can be found in the Appendix.

3.4 Metrics

To evaluate model performance, I measure a variety of metrics:

Accuracy

The accuracy of a classifier is simply the percentage of data points correctly classified. Given that the dataset is imbalanced, this can be a misleading metric on its own.

Recall

Recall gives an idea of the number of false negatives (FNs) a classifier produces. There is very little tolerance for FNs in screening algorithms, since this corresponds to removing a start-up, that would eventually be successful, from the deal flow. A model that produces no FNs will have a recall of 1.

Precision

Precision provides a metric for the number of false positives (FPs) a classifier produces. A FP corresponds to a start-up, that eventually ends up failing, passing through the screening algorithm. This isn't as important as recall since further evaluation by humans can be performed before a decision is made. On the other hand, FNs are directly removed from the deal flow by the screening algorithm, and will have a direct impact on VC profitability.

SHAP Values

SHAP values [36] quantify the contribution that the model features bring towards each prediction. A model for every subset of the feature set i.e. a power set, is trained. From this, the marginal contributions of each feature for each training example can be calculated. The mean absolute SHAP value of a feature can then be calculated to quantify the importance of the feature in the model.

Chapter 4

Results

In this chapter, I present the results of this project. Alongside the results, I provide a discussion to provide physical meaning to the findings. Prior to this, I provide an exploratory data analysis of some key features in the dataset to complement the discussion.

4.1 Exploratory Data Analysis

Distribution by Country

Fig. 4.1 shows the distribution of start-up count and start-up success rates grouped by country in the warm-up period. USA accounts for nearly a third of the start-ups in the warm-up period. This is comparable to 'other' which account for nearly 200 countries. The rate of start-up success appears to vary dramatically by location. China has a start-up success rate nearly double the warm-up window average, whilst countries like India and South Korea are nearly half of the average.

Distribution by Sector

Fig. 4.2 shows the 20 largest sectors by start-up count, along with their corresponding start-up success rates. Interestingly, software, information technology, and internet services sit at top, possibly due to entrepreneurs being inspired by the massive growth of companies like Uber or Airbnb. Some sectors have particularly high success rates, such as Science and Engineering, or Artificial Intelligence, suggesting these are sectors with lots of room for innovation. By contrast, very low success rate sectors like advertising, possibly signal high amounts of competition.

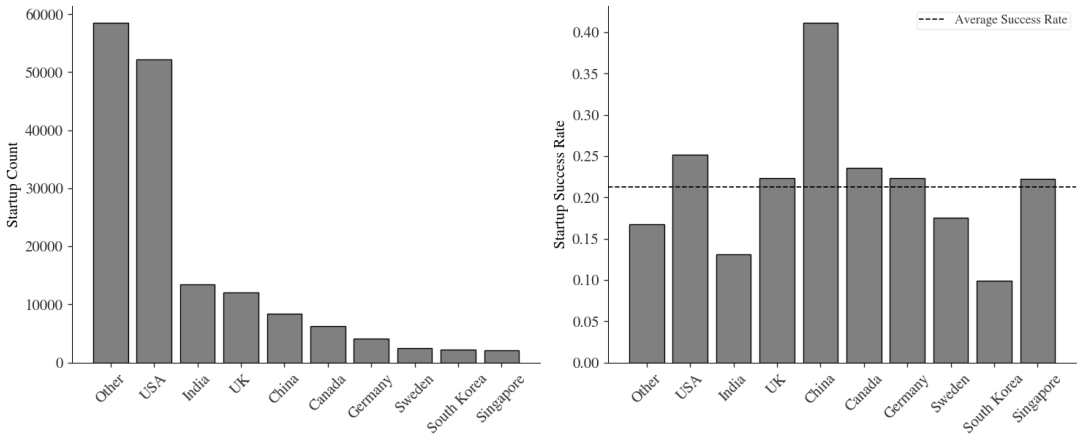

Figure 4.1: A bar chart of the distribution for start-up count (left), and start-up success rates (right), by country in the warm-up window.

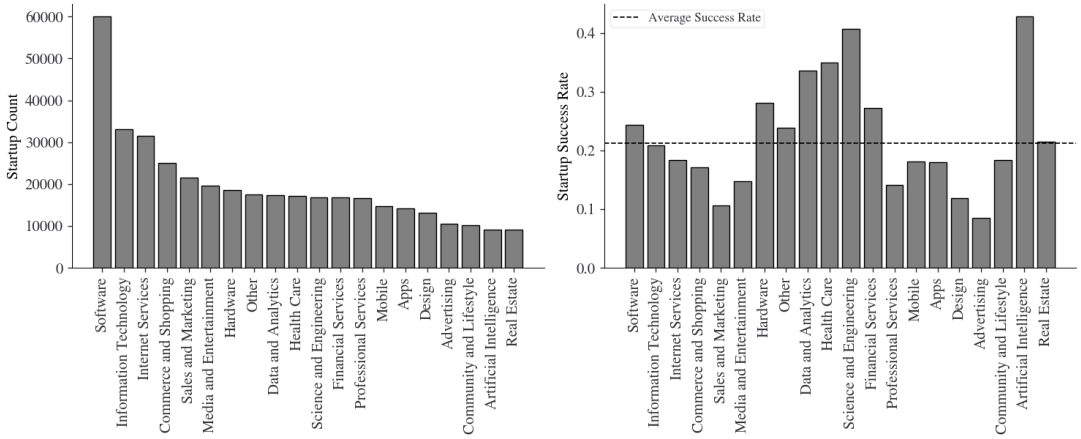

Figure 4.2: A bar chart of the distribution for start-up count (left), and start-up success rates (right), by sector in the warm-up window. Only the top 20 sectors according to start-up count are included.

Syndicate Network Centrality by Investor

Fig. 4.3 shows the normalised count, and box plot of centrality values, for syndicate investors in the warm-up window. VCs and angel investors appear to be the main investors prior to Series A. For all investors, a mean centrality close to zero is seen, suggesting only a minority of investors have the highest centrality values.

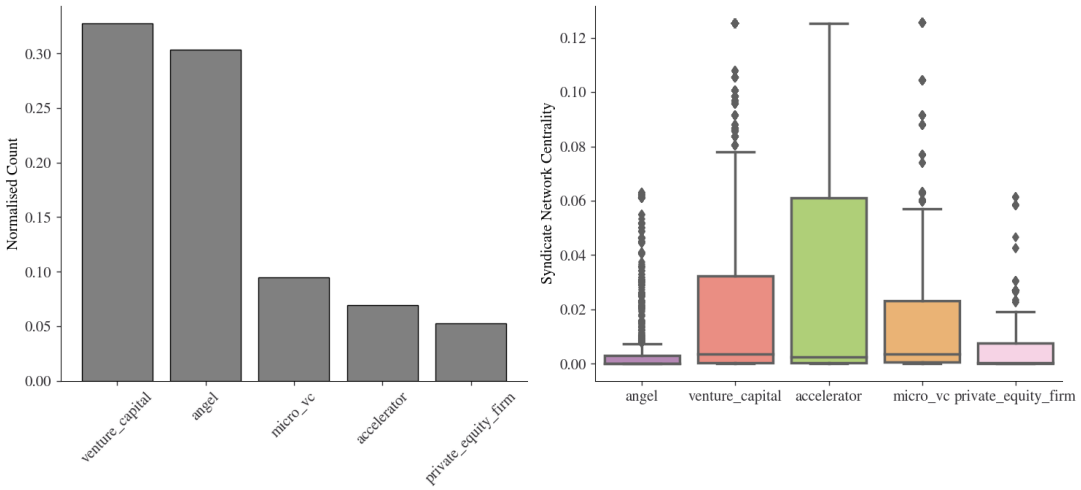

Figure 4.3: On the left, a bar chart of the normalised count of investors in the warm-up window by the investor type. On the right, a box plot of the syndicate network centrality values in the warm-up window for the different investor types.

WWS Network Distribution

Fig. 4.4 shows the probability density curve of the WWS network centrality. Two peaks around 0 and 0.2 are observed. The first peak is a result of the way the closeness centrality algorithm normalises small clusters of disconnected nodes. The main cluster of organisations has a peak density around a centrality of 0.2, with a max value around 0.3.

4.2 Experiment Results

The results obtained from the experiments detailed in Chapter 3 are summarised in Table 4.1. Given the small sample size, standard errors are quoted to 1 significant figure [37][1]. Experiments are separated into four distinct groups based on the feature set used: the baseline, the baseline and WWS network centrality measure (*Baseline + WWS*), the baseline and the syndicate network centrality measure (*Baseline + Syn*), and finally the baseline with both network centrality measures (*Baseline + WWS + Syn*). The difference between a given experiment and the baseline is quoted in grey text, along with the error in the difference (see Appendix A.3), underneath each metric. Noticeably, in all cases I observe accuracy worse than that of a naive classifier i.e. a

[1]Unless the sample size is above 10^4, quoting a standard error with more than 1 significant figure is meaningless because there is also an error in the error [37].

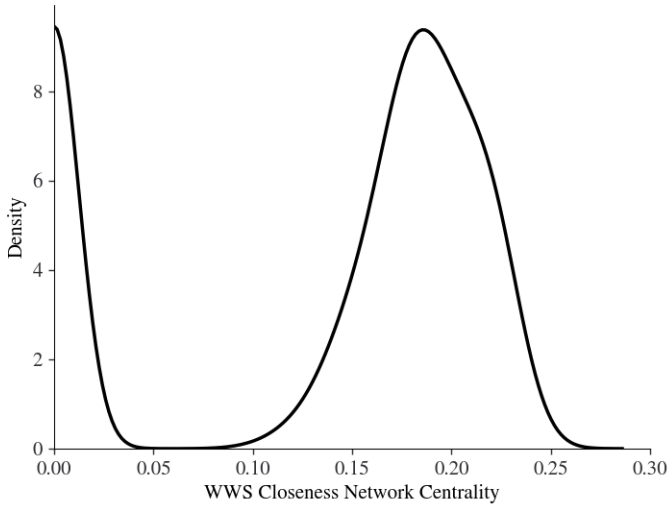

Figure 4.4: A probability density curve of the WWS network closeness centrality values for start-ups in the warm-up window.

classifier which only chooses the dominant class label. A naive classifier would obtain an accuracy of 78.7% since this is the proportion of start-ups labelled unsuccessful. Arroyo et al. [27], as well as Retterath [14], obtain accuracy values in excess of 80% with gradient boosted tree methods. However, given the very early stage funding criteria of the warm-up window, I expected the accuracy to be lower, given the higher risk of investing in a start-up at earlier stages [1]. Furthermore, researchers like Retterath have access to multiple data sources, allowing for a greater number of more enriched features. Only having access to the CB database sets the scope for the remainder of this analysis: determine the importance of the two network features as part of a wider feature set, not obtain the best model metrics in the literature.

Table 4.1: Summary of the results from all XGBOOST models tested in the computational experiment. In the left hand column, baseline represents the base features used throughout all models in the experiment. The inclusion of '+WWS' and/or '+Syn' in the model name represents the addition of the WWS network and/or syndicate network to the model features respectively. The specific centrality measure used for each network feature is indented underneath each model name where applicable. Accuracy, recall, and precision are all stated with a standard error calculated from three repeats of the experiment. A similar procedure was completed for the mean absolute SHAP value averaged over the experiments. The median ranking of the feature relative to all other features has also been provided where applicable.

Models	Metrics			WWS SHAP Values		Syn SHAP Values	
Experiment	Accuracy (%)	Recall (%)	Precision (%)	Mean Absolute SHAP	Median Feature Rank	Mean Absolute SHAP	Median Feature Rank
Baseline	(74.5 ± 0.1)	(70.7 ± 0.4)	(43.9 ± 0.1)	-	-	-	-
Baseline + WWS							
Closeness	(75.0 ± 0.1) (0.5 ± 0.1)	(70.9 ± 0.3) (0.2 ± 0.5)	(44.5 ± 0.1) (0.6 ± 0.1)	(0.098 ± 0.002)	8	-	-
Betweenness	(75.0 ± 0.1) (0.5 ± 0.1)	(71.0 ± 0.5) (0.3 ± 0.6)	(44.4 ± 0.1) (0.5 ± 0.1)	(0.097 ± 0.003)	8	-	-
Degree	(75.0 ± 0.1) (0.5 ± 0.1)	(70.5 ± 0.3) (-0.2 ± 0.5)	(44.5 ± 0.1) (0.6 ± 0.1)	(0.092 ± 0.002)	8	-	-
Eigenvector	(74.9 ± 0.1) (0.4 ± 0.1)	(71.0 ± 0.3) (0.3 ± 0.5)	(44.4 ± 0.1) (0.5 ± 0.1)	(0.100 ± 0.002)	8	-	-
Baseline + Syn							
Closeness	(74.81 ± 0.05) (0.3 ± 0.1)	(70.3 ± 0.4) (-0.4 ± 0.6)	(44.16 ± 0.06) (0.3 ± 0.1)	-	-	(0.066 ± 0.003)	13
Betweenness	(74.72 ± 0.04) (0.2 ± 0.1)	(70.5 ± 0.4) (-0.2 ± 0.6)	(44.06 ± 0.03) (0.2 ± 0.1)	-	-	(0.063 ± 0.003)	15
Degree	(74.67 ± 0.04) (0.2 ± 0.1)	(70.6 ± 0.4) (-0.1 ± 0.6)	(44.01 ± 0.02) (0.1 ± 0.1)	-	-	(0.070 ± 0.002)	12
Eigenvector	(74.93 ± 0.05) (0.4 ± 0.1)	(70.0 ± 0.2) (-0.7 ± 0.5)	(44.30 ± 0.08) (0.4 ± 0.1)	-	-	(0.067 ± 0.003)	12
Baseline + WWS + Syn							
Closeness + Degree	(74.96 ± 0.04) (0.5 ± 0.1)	(71.0 ± 0.4) (0.3 ± 0.6)	(44.41 ± 0.07) (0.5 ± 0.1)	(0.101 ± 0.002)	7	(0.070 ± 0.005)	14
Closeness + Eigenvector	(75.10 ± 0.06) (0.6 ± 0.1)	(70.8 ± 0.4) (0.1 ± 0.6)	(44.58 ± 0.05) (0.7 ± 0.1)	(0.098 ± 0.003)	8	(0.070 ± 0.006)	13

4.2.1 Baseline and WWS Network

For all centrality measures in the *Baseline + WWS* experiment, I observe roughly a 0.5% increase in accuracy and precision. It is unclear how the WWS network centrality measure affects recall given the relatively large standard errors. I see a tight distribution in the median feature rank for all centrality measures, highlighting that the measures provide more or less the same amount of importance. In fact, this trend was also observed by Bonaventura et al. [16], who found closeness, betweenness, and degree to be closely correlated in the WWS network. Mean absolute SHAP values show that eigenvector has a slight edge over degree given that they both disagree within a standard error of each other. Physically, this means that a start-up hiring expertise many companies is important (degree), but more so when they come from a company that also hires expertise from many other companies (eigenvector). Although, in practice, the difference is very small given the similarity among the model metrics. Nonetheless, the experimental results have proved Hypothesis 1b for accuracy and precision; however, the feature's contribution towards recall is unclear.

4.2.2 Baseline and Syndicate Network

Similar to *Baseline + WWS*, I observe increases in accuracy and precision for the majority of syndicate network centrality measures. However, some outliers in the feature set exist; no statistically significant improvement in precision for degree centrality is observed. Furthermore, I observe a decrease in recall for eigenvector centrality within a standard error. One possible explanation for this is the *curse of dimensionality*: the extra dimension of the feature set outweighs the benefit of adding the feature in the first place. Comparing holistically to *Baseline + WWS*, there are two key observations. First, any observed increases of model metrics are of a lower magnitude. This suggests that the WWS network centrality is a more important feature to the classifier than the syndicate network centrality. There is a possibility that, if the syndicate network had significantly more missing values than the WWS network, the former could appear less important. However, both network features have very similar percentages of missing values at around 89%. Second, a greater variation across the centrality measures is observed. The implication is that certain centrality measures are more informative than others. This is further backed up by the variation in median feature rank among the four centrality measures.

Physically, betweenness represents an investors ability to act as an intermediary be-

tween other investors, whilst closeness serves as a likelihood measure that the investor has access to more expertise, knowledge, and social capital. Finally, degree represents an investor who has access to a wider range of knowledge and expertise from co-investors, whilst eigenvector also takes into account the range of knowledge and expertise of the co-investors. Median feature ranks show that start-ups benefit the least from investors who act as intermediaries, whilst the other 3 centrality measures serve as good proxies for the knowledge and expertise they can obtain from their investor, with eigenvector and degree showing a slight edge in median feature rank. In fact, the results observed are in agreement with Hochberg et al. [19], who observe eigenvector and degree to have the highest correlation with a successful outcome. Hypothesis 2b is true to a lesser extent than Hypothesis 1b, showing to be true for accuracy and precision, except for degree centrality. Similarly, recall is unclear, apart from eigenvector centrality where a decrease was observed.

4.2.3 Baseline, WWS Network, and Syndicate Network

Informed by the previous results, I chose the best performing centrality measure from each experiment to be used in a model that incorporates both network features. Ideally, every combination would be tested; however, there were limits on computational resources. For *Baseline + WWS*, I experimentally observed a low amount of variation between the different centrality measures in all metrics measured. This observation, coupled with Bonaventura et al. [16] ultimately using closeness centrality in their main publication, I selected the same centrality measure. For the syndicate network, median feature rank suggests either degree or eigenvector. Paired with the findings of Hochberg et al., I chose to experiment with both degree and eigenvector.

Of both the models tested, I observe similar performance. Eigenvector centrality (*Closeness + Eigenvector*) has a slight edge over degree (*Closeness + Degree*) centrality for the syndicate network. Comparing the two, *Closeness + Eigenvector* provides a greater accuracy and precision than *Closeness + Degree* within a standard error. Once again, it is unclear which model provides better values of recall since they agree within their standard errors. Overall, the best performing model incorporates closeness centrality for the WWS network, and eigenvector centrality for the syndicate network. Fig. 4.1 summarises the top 15 features of the *Closeness + Eigenvector* model.

The left hand side of Fig. 4.1 plots the top 15 features in descending order of mean absolute SHAP value. The right hand side of Fig. 4.1 provides an overview of how

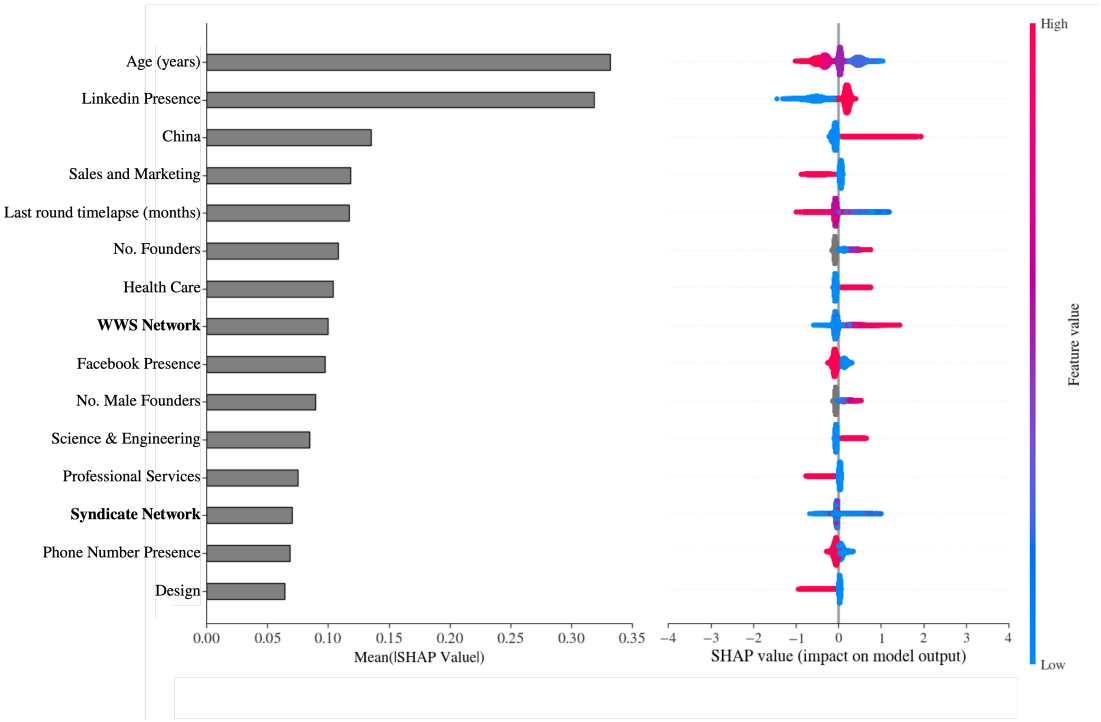

Figure 4.5: On the left, a bar chart of the most importance model features, in descending order of their mean absolute SHAP value. On the right, a plot of the distribution of SHAP values for each feature. Vertical height in this plot represents a higher density of data points.

SHAP values vary with feature value. Greater vertical spread represents higher count of data points. Upon inspection, some of the most important features are time related. The age of the start-up in years appears to be negatively correlated with feature importance, with a similar trend seen for the time elapsed between the last funding round and the start of the simulation window. This suggests that older start-ups that haven't been funded recently are struggling to get funding beyond seed, and serves as a proxy for a struggling start-up. Trends can also be seen in sector, where those with low start-up success rate in Fig. 4.2 are negatively correlated with success, whilst the opposite is true for sectors with a high success rate. Economically, this could correspond to markets that are saturated and competitive (sales & marketing and design), or markets that have plenty of room for innovation (science & engineering and health care). Features such as the number of male founders also have a positively correlated trend, raising questions surrounding gender biases in the start-up industry. Social media presence also appears to be important, with particular emphasise on professional social networking (LinkedIn). Moreover, China appears to be a good indicator due to the high

success rate observed in Fig 4.1. Finally, SHAP value appears to be positively correlated with WWS network centrality, whereas the syndicate network centrality measure is unclear in this representation and requires further investigation.

4.3 Network Features SHAP Analysis

4.3.1 WWS Network

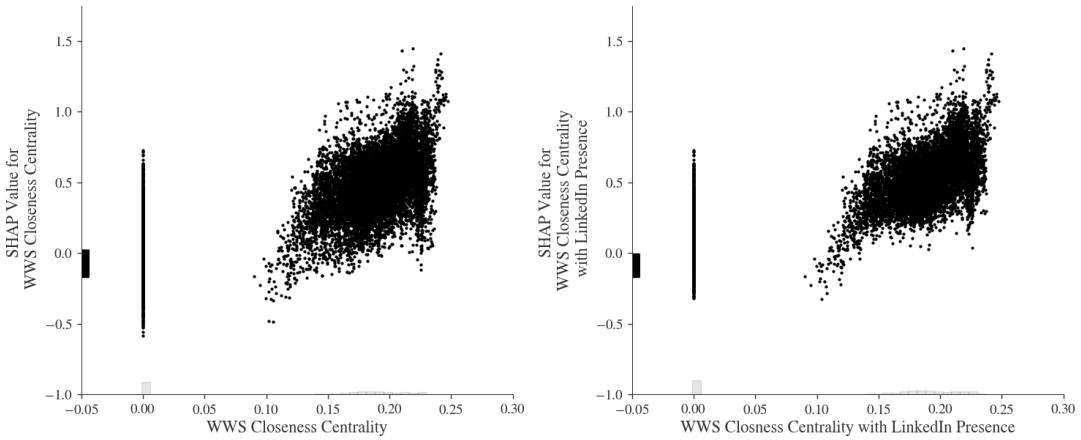

Figure 4.6: On the left, a plot of SHAP value against WWS closeness centrality value. On the right, the same plot, but conditioned on the start-up having LinkedIn.

The left hand side of Fig. 4.6 shows the variation in SHAP value against the WWS network closeness centrality measure. Within this plot, the very left hand side cluster of data points represents start-ups with a missing centrality measure (i.e. the start-up did not have the data for it). In the right-hand cluster of data, a positive correlation can be seen between WWS network centrality and SHAP value. Hence, this proves Hypothesis 1a.

A vertical spread in the correlation shows there are other features interacting with the SHAP value distribution, one of which was found to be LinkedIn presence. On the right hand side of Fig 4.2, I plot the same graph, but conditioned on the start-up having LinkedIn presence. Visually, data points towards the bottom of the positive correlation are no longer present. This suggests that professional social media presence increases the start-ups exposure to professionals in the industry. This could put the start-up in a better position to hire superior talent and expertise, from which the start-up will benefit.

4.3.2 Syndicate Network

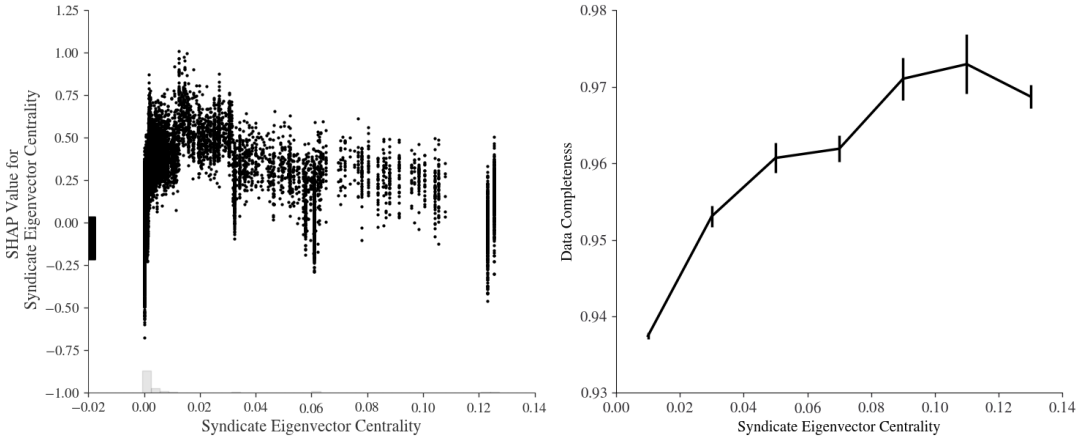

Figure 4.7: On the left, a plot of SHAP value against syndicate network eigenvector centrality value. On the right, a plot of data completeness against syndicate eigenvector centrality, with values averaged in bin sizes of 0.02. Vertical black lines represent the standard error in the mean.

The left hand side of Fig. 4.7 shows the variation in SHAP value against syndicate network eigenvector centrality. Intuitively, one might expect a correlation much like the one seen with the WWS network. The higher the syndicate network eigenvector centrality, the more access the investor has to better knowledge and expertise from other well connected investors. The instinctive assumption is that this will benefit the start-up more, thereby increasing their likelihood of success. However, I do not observe this in the experimental results. The benefit of having of a well connected investor, prior to Series A, increases to a peak value, then beyond this peak, a counter-intuitive negative correlation is observed.

Given this counter-intuitive result, I investigate the negative correlation further. The majority of the entries in the dataset are not fully complete, with empty values for some of the features. The right hand side of Fig. 4.7 plots the average completeness of a start-up against investor centrality. For a given start-up, I define completeness as the ratio of non-empty features to the total number of features in the dataset. The average was computed in centrality bin sizes of 0.02. The plot shows a trend that start-ups associated with investors of high centrality in the syndicate network have a more complete dataset. Any biases in the data have implications for the observed trends in SHAP values since it is a local measure. For example, in the absence of full data, XGBOOST will see a large amount of value in relatively small differences of investor

centrality. However, should the algorithm now have access to several more features, which is the case for start-ups associated with high centrality investors, the value of investor centrality could be artificially decreased. To investigate this further, I control for data completeness, following the same training procedure in Sec. 3.3.2. Any features containing empty values were removed from the dataset. This left 63 features, along with the syndicate network eigenvector centrality and dependent variable.

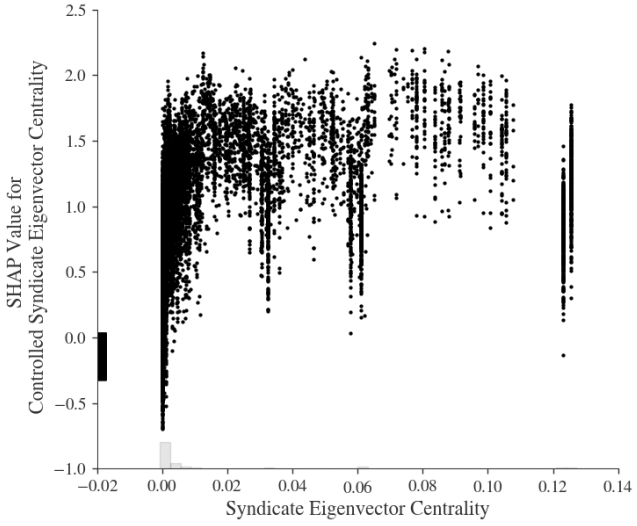

Figure 4.8: A plot of SHAP value against syndicate network eigenvector centrality value, but controlling for data completeness.

The plot of SHAP value against syndicate network eigenvector centrality, controlling for completeness bias, is shown in Fig. 4.8. Controlling for data completeness removed the counter-intuitive negative correlation observed in Fig. 4.7. However, even with this control, a plateau after the a sharp initial peak is observed. Therefore, Hypothesis 2a is only true up to a centrality of roughly 0.02. In an attempt to spot any trends amongst different types of investors in the warm-up window, I split Fig. 4.8 into the five biggest seed investors in the dataset (see Fig. 4.9). Visually, each type of investor more or less contain the same plateau trend. Therefore, the observed trend is potentially an artefact of investments prior to Series A, not the type of investor. However, such work would require investigating many different selection criteria for the warm-up window, a task beyond the scope of this project.

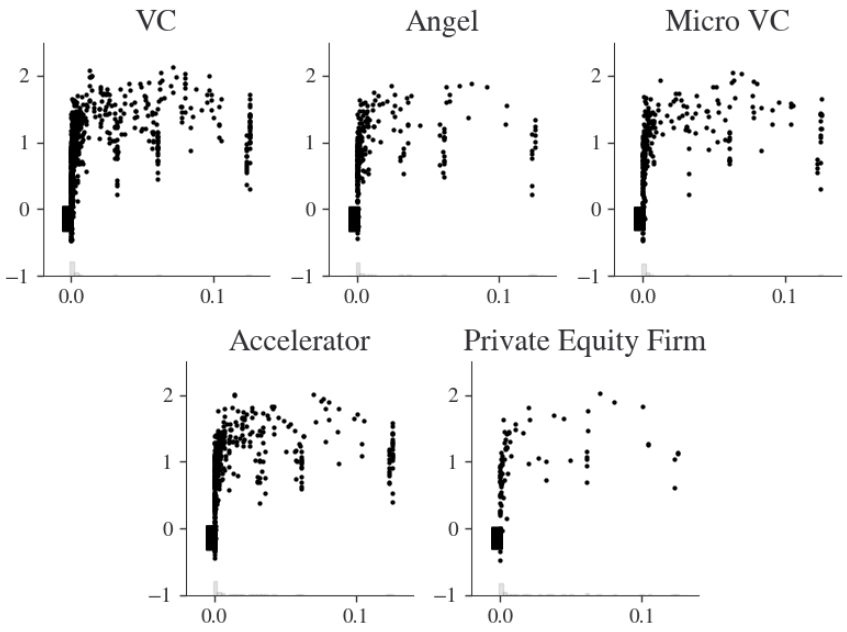

Figure 4.9: Plots of SHAP value against syndicate network eigenvector centrality value, controlling for data completeness, and grouped by the investor type.

Chapter 5

Conclusions

5.1 Limitations and Future Work

The results presented in Sec 4.3.2 have provided an excellent foundation for further work. The observed plateau seems to suggest that there is a limit to the benefits a better connected investor can provide to a start-up. The founders of start-ups normally accept a 10-14% discount [38] to associate with high status investors; however, the findings appear to suggest that, beyond a certain point, the increased cost is not correlated with a higher likelihood of a successful outcome. This opens up an avenue of research questions, such as: do high-status investors in the seed stage increase the likelihood of start-up success? Answering such a question would require an in-depth analysis, beyond the scope of this project, and would be of particular interest to the founders of a start-up. Such an analysis could also uncover how the median feature rank varies as later stages of funding are considered. Hochberg et al. [19] reported that syndicate centrality was able to explain the variation in survival rates less as they considered later funding rounds. It would be interesting to see if their finding would be reciprocated.

Although beyond the scope, incorporating multiple databases to enrich and increase the number of features would have been ideal. Without data limitation, one of the top XGBOOST classifiers for predicting start-up success [14] could have been created as our baseline. This would have provided more interesting comparisons between the models used in this project. Given the expense of these databases, it was not possible to obtain such baseline results. Furthermore, researchers with access to additional databases could enrich the network features, as well as other features in the dataset, to investigate how to maximise the importance of the network-based features.

For experimental results, I prioritised computational budgets on extensive hyper-

parameter searching, along with several repeats such that statistically meaningful comparisons could be made between models. As a consequence, this prevented the ability to experiment with other ML methods. Ideally, a range of models from simple decision trees, to deep neural networks would have been tested and compared. It also would have been interesting to see if the analysis of network-based features varies by ML method. Given the importance of recall for a screening tool in venture capital, it is unfortunate to see such relatively high standard errors for all recall metrics. In future, further runs of the experiment would reduce the standard error of recall such that comparisons can be made. Although, my results emphasise that any massive leaps in recall (and any other model metric for that matter) will likely not be obtained from one or two additional features. Moreover, it raises the questions of the practicality of implementing such features in the first place given the computational expense of calculating the metric. One could turn to the CB rank of an investor should computational budgets be a problem. CB provides a rank [39] for each investor based off of news articles, community engagement, funding rounds etc. The marginal contribution of an investors CB rank could be explored. However, this would require access to historical values of the rank, which cannot be obtained with .csv export of the CB database.

An early-stage VC might argue that the methodology presented would not be the most realistic in a real world setting. A VC using quantitative sourcing methods is constantly scraping websites, thus the deal flow is continuous. The ideal ML screening algorithm will also be continuous, possibly utilising sliding sliding time windows, not imposing an arbitrary simulation date. However, such methods are extremely computationally expensive and out of reach for this project. Nonetheless, the analysis of the network-based features is still a valuable contribution which could be utilised in a VC setting.

Finally, there has been a recent trend from entrepreneurs bypassing angel and seed investments in favour of backing from VCs [7]. This is mainly due to the decreasing costs of creating a start-up. Third party services such as Amazon Web Services and shared work spaces have allowed founders to develop their product or service with much less funding compared to a decade ago. Over the next 5 to 10 years, it would be interesting to see how the importance of the syndicate network for investors prior to Series A changes. Should it lose importance over time, it would raise questions about the practicality of the feature given the computational expense.

5.2 Concluding Remarks

In this dissertation, I used the Crunchbase database to investigate two network-based features from the literature: an investor syndicate network, and a professional relations amongst companies network. The experimental results showed the latter to be more important in terms of model metrics and SHAP analysis than the former. The professional relations network centrality values showed to be positively correlated with start-up success. Conversely, the syndicate network centrality values displayed a rapid positive correlation into a plateau, suggesting there is little benefit associated with upper echelon investors at the early stage of the start-up funding life cycle.

This work directly targeted the screening process of an early-stage VC, a key stage of the investment process which is a good target for automation. Recall is of particular importance at this stage of the process, since missing out on the next Google or Facebook would directly affect the profitability of the VC. My experiments with recall had large standard errors, thus any statistically meaningful comparisons on recall improvement could not be drawn. Further repeats of the experiment in the future would reduce the standard error allowing for more meaningful comparisons. The main direct contribution of this work was uncovering the underlying value of the two network-based features. It is hoped an even larger indirect contribution will be made by promoting further research in a field which will be rapidly evolving over the next 5 years [12].

Bibliography

[1] John C. Ruhnka and John E. Young. Some hypotheses about risk in venture capital investing. *Journal of Business Venturing*, 1991.

[2] Paul A. Gompers, Josh Lerner, Margaret M. Blair, and Thomas Hellmann. What drives venture capital fundraising? *Brookings Papers on Economic Activity. Microeconomics*, 1998.

[3] Andrew L Zacharakis and Dean A Shepherd. The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, 2001.

[4] Nikolaus Franke, Marc Gruber, Dietmar Harhoff, and Joachim Henkel. What you are is what you like—similarity biases in venture capitalists' evaluations of start-up teams. *Journal of Business Venturing*, 2006.

[5] Andrew L. Zacharakis and G. Dale Meyer. A lack of insight: Do venture capitalists really understand their own decision process? *Journal of Business Venturing*, 1998.

[6] Daniel Stangler and Paul Kedrosky. Exploring firm formation: Why is the number of new firms constant? *SSRN Electronic Journal*, 2012.

[7] PitchBook. Annual european venture report, 2019.

[8] Paul A. Gompers, Will Gornall, Steven N. Kaplan, and Ilya A. Strebulaev. How do venture capitalists make decisions? *Journal of Financial Economics*, 2020.

[9] S. Benson Edwin Raj and A. Annie Portia. Analysis on credit card fraud detection methods. 2011.

[10] Kristian Bondo Hansen. The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data and Society*, 2020.

[11] Citi. Bank of the Future: the ABCs of digital disruption in finance, 2018.

[12] Meghan Rimol and Katie Costello. Gartner says tech investors will prioritize data science and artificial intelligence above "gut feel" for investment decisions by 2025. URL `https://www.gartner.com/en/newsroom/press-releases/2021-03-10-gartner-says-tech-investors-will-prioritize-data-science-and-artificial-intelligence-above-gut-feel-for-investment-decisions-by-20250`. (Accessed as of 17/08/2021).

[13] Andre Retterath and Reiner Braun. Benchmarking venture capital databases. *SSRN Electronic Journal*, 2020.

[14] Andre Retterath. Human versus computer: Benchmarking venture capitalists and machine learning algorithms for investment screening. *SSRN Electronic Journal*, 2020.

[15] Vance H. Fried and Robert D. Hisrich. Toward a model of venture capital investment decision making. *Financial Management*, 1994.

[16] Moreno Bonaventura, Valerio Ciotti, Pietro Panzarasa, Silvia Liverani, Lucas Lacasa, and Vito Latora. Predicting success in the worldwide start-up network. *Scientific Reports*, 2020.

[17] Steven N. Kaplan and Per Strömberg. Characteristics, contracts, and actions: Evidence from venture capitalist analyses. *Journal of Finance*, 2004.

[18] Thomas Hellmann and Manju Puri. Venture capital and the professionalization of start-up firms: Empirical evidence. *Journal of Finance*, 2002.

[19] Yael V. Hochberg, Alexander Ljungqvist, and Yang Lu. Whom you know matters: Venture capital networks and investment performance. *Journal of Finance*, 2007.

[20] Andrew Lockett and Michael Wright. The syndication of private equity: Evidence from the uk. *Venture Capital*, 1999.

[21] Joshua Lerner. The syndication of venture capital investments. *Financial Management*, 1994.

[22] James A. Brander, Raphael Amit, and Werner Antweiler. Venture-capital syndication: Improved venture selection vs. the value-added hypothesis. *Journal of Economics and Management Strategy*, 2002.

[23] Miguel Meuleman, Mike Wright, Sophie Manigart, and Andy Lockett. Private equity syndication: Agency costs, reputation and collaboration. *Journal of Business Finance and Accounting*, 2009.

[24] Olav Sorenson and Toby E. Stuart. Syndication networks and the spatial distribution of venture capital investments. *American Journal of Sociology*, 2001.

[25] William L. Megginson and Kathleen A. Weiss. Venture capitalist certification in initial public offerings. *The Journal of Finance*, 1991.

[26] Paula Väätänen. Angel investor syndication: Syndicate formation and networks in finland. *Master's Thesis for Jyväskylä University*, 2021.

[27] Javier Arroyo, Francesco Corea, Guillermo Jimenez-Diaz, and Juan A. Recio-Garcia. Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, 2019.

[28] Gerard J. Tellis. 2016 startup index of nations. 2016. URL `https://www.marshall.usc.edu/sites/default/files/Unicorn-Index-Report-GT17.pdf`.

[29] Linton C. Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1978.

[30] Katherine Fraust and Stanley Wasserman. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[31] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 1977.

[32] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 2001.

[33] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 1987.

[34] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. 2016.

[35] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. 2019.

[36] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. 2017.

[37] Ifan G. Hughes and Thomas P. A. Hase. *Measurements and their Uncertainties*. Oxford University Press, 2010.

[38] David H. Hsu. What do entrepreneurs pay for venture capital affiliation? *Journal of Finance*, 2004.

[39] Denise Stephan. What is crunchbase rank and trend score? URL `https://about.crunchbase.com/blog/crunchbase-rank-trend-score/`. (Accessed as of 17/08/2021).

Appendix A

Additional Data

A.1 Degree Abbreviations

Bachelor's Degree

Degree	Abbreviations
Bachelor of Arts	ba
Bachelor of Arts	ab
Bachelor of Arts	barts
Bachelor of Science And Arts	bsa
Bachelor of Accountancy	bacy
Bachelor of Accounting	bacc
Bachelor of Animal and Veterinary Bioscience	banvetbiosc
Bachelor of Applied Science	bappsc
Bachelor of Applied Science	basc
Bachelor of Architecture	barch
Bachelor of Business Administration	bba
Bachelor of Civil Engineering	bce
Bachelor of Commerce	bcomm
Bachelor of Commerce	bcom
Bachelor of Communications	bcomm
Bachelor of Computer Application	bca
Bachelor of Dental Hygiene	bdh
Bachelor of Dental Medicine	bdm
Bachelor of Dental Science	bdsc
Bachelor of Dental Surgery	bds
Bachelor of Dental Surgery	bchd
Bachelor of Dentistry	bdent
Bachelor of Design	bdes
Bachelor of Design Computing	bdescomp
Bachelor of Design in Architecture	bdesarch
Bachelor of Education	bed
Bachelor of Engineering	beng
Bachelor of Engineering	be
Bachelor of Electronic Commerce	bec

Degree	Abbreviations
Bachelor of Electronic Commerce	be-com
Bachelor of Electrical Engineering	bee
Bachelor of Fine Arts	bfa
Bachelor of Health Sciences	bhlthsci
Bachelor of Information Technology	bit
Bachelor of International and Global Studies	bigs
Bachelor of Law	llb
Bachelor of Liberal Arts and Sciences	blas
Bachelor of Library Science	blib
Bachelor of Library Science	bls
Bachelor of Literature	blit
Bachelor of Mathematics	bmath
Bachelor of Mechanical Engineering	bme
Bachelor of Medical Science	bmedsc
Bachelor of Medicine	mb
Bachelor of Music	bm
Bachelor of Music Studies	bmusstudies
Bachelor of Nursing	bn
Bachelor of Pharmacy	bpharm
Bachelor of Philosophy	bph
Bachelor of Policing	bpol
Bachelor of Policing (Investigations)	bpoli
Bachelor of Professional Studies	bps
Bachelor of Resource Economics	bresec
Bachelor of Science	bs
Bachelor of Science	bsc
Bachelor of Science in Dental Hygiene	bsdh
Bachelor of Science in Environmental and Occupational Health	bseoh
Bachelor of Science in Nursing	bsn
Bachelor of Socio-Legal Studies	bsls
Bachelor of Surgery	bs
Bachelor of Technology	btech
Bachelor of Veterinary Science	bvsc
Bachelor of Visual Arts	bva

Master's Degree

Degree	Abbreviations
Master of Architecture (Professional Degree) or Master of Science in Architecture (Research Degree)	march
Master of Architecture (Professional Degree) or Master of Science in Architecture (Research Degree)	ms
Master of Arts	ma
Master of Arts	am
Master of Business Administration	mba
Master of Chemistry	mchem
Master of Commerce	mcom
Master of Computer Application	mca
Master of Divinity	mdiv
Master of Education (Master of Science in Education)	med
Master of Education (Master of Science in Education)	msed
Master of Emergency Management	mem
Master of Emergency and Disaster Management	medm

Degree	Abbreviations
Master of Engineering	me
Master of Engineering	meng
Master of Fine Arts	mfa
Master of Health or Healthcare Management	mschm
Master of Health or Healthcare Management	mhm
Master of Health Infromatics	mschi
Master of Health Infromatics	mhi
Master of International Affairs	mia
Master of International Studies	mis
Master of Laws	llm
Master of Library Science	mls
Master of Liberal Arts	mla
Master of Library and Information Science	mlis
Master of Music	mm
Master of Professional Studies	mps
Master of Public Administration	mpa
Master of Public Health	mph
Master of Science	ms
Master of Science	msc
Master of Science in Information	msi
Master of Social Work	msw
Master of Strategic Foresight	msf
Master of Sustainable Energy and Environmental Management	mseem
Master of Technology	mtech
Master of Technology Managment	mtm
Master of Theology	thm
Master of Philosophy	mphil
Master of Physics	mphys
Master of Mathematics	mmath
Master of Science	msci

Doctorate

Degree	Abbreviations
Doctor of Acupuncture	dac
Doctor of Audiology	aud
Doctor of Biblical Studies	dbs
Doctor of Biblical Studies	db
Doctor of Chiropractic	dc
Doctor of Dental Surgery	dds
Doctor of Divinity	dd
Doctor of Education	edd
Doctor of Jurisprudence (Juris Doctor)	jd
Doctor of Immortality	imd
Doctor of Law and Policy	lpd
Doctor of Law and Policy	dlp
Doctor of Medical Dentistry	dmd
Doctor of Medicine	md
Doctor of Ministry	dmin
Doctor of Metaphysics	drmph
Doctor of Musical Arts	dma

Degree	Abbreviations
Doctor of Naturopathy	nd
Doctor of Nursing Practice	dnp
Doctor of Optometry	od
Doctor of Osteopathy	do
Doctor of Pharmacy	pharmd
Doctor of Philosophy	phd
Doctor of Philosophy	dphil
Doctor of Philosophy	dph
Doctor of Physical Therapy	dpt
Doctor of Practical Theology	dpt
Doctor of Psychology	psyd
Doctor of Public Health	drph
Doctor of Religious Sciences	drscrel
Doctor of Religious Sciences	drs
Doctor of Science	dsc
Doctor of Science	scd
Doctor of Theology	dth
Doctor of Theology	thd
Doctor of Veterinary Medicine	dvm

A.2 Hyper-parameter Search Space

Below are the hyper-parameters search using the Optuna hyper-parameter optimiser package for python.

```
param = {
    "verbosity": 0,
    "objective": "binary:logistic",
    "eval_metric": "auc",
    "scale_pos_weight": 3.7,
    "booster": "gbtree",
    "lambda": trial.suggest_float("lambda", 1e-8, 1.0, log=True),
    "alpha": trial.suggest_float("alpha", 1e-8, 1.0, log=True),
    "subsample": trial.suggest_float("subsample", 0.2, 1.0),
    "colsample_bytree": trial.suggest_float("colsample_bytree", 0.2, 1.0),
    "colsample_bylevel": trial.suggest_float("colsample_bylevel", 0.2, 1.0),
    "min_child_weight": trial.suggest_float("min_child_weight", 0.01, 20.0, log=True),
    "max_depth": trial.suggest_int("max_depth", 1, 9),
    "eta": trial.suggest_float("eta", 1e-8, 1.0, log=True),
    "gamma": trial.suggest_float("gamma", 1e-8, 1.0, log=True),
    "grow_policy": trial.suggest_categorical("grow_policy", ["depthwise", "lossguide"])
    }
```

A.3 Uncertainties

The uncertainty about the mean (standard error) for a distribution, A, is given by

$$\alpha_A = \frac{\sigma_{n-1}}{\sqrt{n}} \tag{A.1}$$

where σ is the (sample) standard deviation and n is the sample size. The difference of two means, $Z = A - B$, also has a standard error given by

$$\alpha_Z = \sqrt{(\alpha_A)^2 + (\alpha_B)^2} \tag{A.2}$$