**RQ: What factors affect startup funding at different funding rounds?**

**0. Introduction**

**1. Theoretical Background**

* Entrepreneurship
* Venture capital
* Signalling theory
* Development over time

**2. Model Selection**

**3. Data Sources**

**3.1. Data Collection**

Online data collection can be far more efficient than historical methods of collecting data for private company investment analyses because online data sources often have interfaces that allow them to be digested automatically and at scale. Data collected from these online data sources can be publically-available or private and require authentication. Publically-available data is far easier to work with because it can often be analysed without any user intervention. However, private data is potentially richer and more useful for prediction.

Public pre-formatted data dumps are often available for research purposes, which allows researchers who do not have a computer science background to gain access to these datasets. An alternative to a pre-formatted data dump is a dataset crawler that uses the source's application programming interface (API) to traverse through the dataset, collecting publically-available data. The advantages of such a crawler is that it can selectively collect data from nodes with specific attributes, or collect a random sample (as in \cite{jang2015}), or traverse the data source indefinitely, updating entries as new data becomes available.

**3.2. Basic Company Data**

CrunchBase is an open online database of information about startups, investors, media coverage and trends, focusing on high-tech industry in the United States. It relies on its online community to edit most pages. CrunchBase is a very comprehensive database, with almost complete coverage of startups and investors in the Internet sector, including the relationships between them \cite{alexy2012}. In May 2014, Zhao et al. collected 62,926 investment events between 7,706 VCs and 18,026 startups from 1987 to 2014 on CrunchBase \cite{zhao2015}. However, it has been noted that the CrunchBase corpus is sparse with many missing attributes \cite{xiang2012, zhao2015}.

CrunchBase has two provisions to prevent and remediate inaccurate crowd-sourced entries \cite{crunchbase2014}. First, all users are require to authenticate their CrunchBase accounts with a social media account which allows CrunchBase to verify a user's true identity. Second, every change goes through a machine review, which flags significant or questionable updates for moderation. That aside, further outlier removal will probably be required for analyses using CrunchBase data.

AngelList is a promising new source of startup data, combining the functionality of an equity crowdfunding platform, a social networking site and an online startup database.

As an equity crowdfunding platform, users create profiles for their startups on AngelList, and use the platform to attract investment. Investors use the platform to identify investment opportunities and can invest directly through AngelList, often alongside other investors in investment syndicates.

AngelList is also an online startup database. It has a data-sharing agreement with CrunchBase (which results in significant overlap between the two sources). Importantly, it tracks 'startup roles' (e.g. founders, investors, employees) with a creation time, start time and end time.

Despite being almost entirely absent from the academic literature, AngelList offers substantial data on startups, entrepreneurs and investors which has the potential for some novel insights. For example, Britz et al. \cite{britz2013} analysed the formation of relationships on AngelList (looking at edge-creation time-stamps) to perform a longitudinal study of community growth and development.

**3.3. Financial Data**

**3.4. Social Networks & Media**

LinkedIn has been used in many online social network studies of entrepreneurship because it is a social network primarily used for professional networking. Unfortunately, as of May 2015, the LinkedIn API no longer allows access to authenticated users' connection data or company data \cite{trachtenberg2015}, making it virtually impossible to use this site for social network analysis, without resorting to semi-automatic HTML parsing techniques (which is against the Terms of Service).

Facebook is a friendship-based online social network. Like LinkedIn, it has also been used quite commonly for entrepreneurial online social network studies \cite{song2012, gloor2013}. In contrast with LinkedIn, Facebook profiles involve significant personal / non-business information. Entrepreneurs' networks on Facebook tend to consist of stronger, more personal ties than entrepreneurs' networks on LinkedIn \cite{gloor2013}.

Instagram is a media-sharing social network particularly popular amongst young people \cite{jang2015}. Accounts are public by default, unless users elect to create a private account. Despite the popularity of Instagram, there is little scholarly work on it, especially in the field of entrepreneurship. However, because of its younger demographic it may provide insights into the extent to which the social behaviour of entrepreneurs changes over generations.

Twitter is a social networking site and micro-blogging site that is often used by entrepreneurs to promote their personal and business brands and rapidly share news and opportunities. Users can send and read public messages (called tweets) of 140 character length. Twitter is a directed network where users are able to follow other users without gaining their permission to do so. Song and Vinig \cite{song2012} found that Twitter network size had a negative relationship with startup survival, perhaps because some entrepreneurs expend too much effort managing this network.

AngelList also acts as a social network. Users create profiles for themselves on AngelList which they populate with information about their career and interests, like LinkedIn. Like Twitter, users can follow other users without gaining their permission to do so. These relationships are time-stamped and accessible through the API.

**3.5. Patents & Intellectual Property**

**3.6. Summary**

Online data sources typically capture a wide spectrum of businesses and entrepreneurs with varying data quality. For example, CrunchBase, an online startup database, includes large companies like Google and Ebay as well as small startups founded very recently. It is important to filter entities to some extent before further analyses are performed. One possible strategy is to only analyse companies that have receive venture funding, are in the IT or internet sectors, and are based in the United States, because this is likely to be the largest grouping (as in \cite{alexy2012}).

Appropriate selection of data sources is critical in startup investment analyses because different data sources provide insights into different actors, relationships and attributes \cite{formsma2012, song2012}.

**4. Algorithm Selection**

The key objective of machine learning algorithm selection is to find algorithms that make assumptions that are consistent with the structure of the data (e.g. tolerance to missing values, mixed feature types, imbalanced classes) and suit the constraints of the application problem (e.g. time available, incremental learning, interpretability).

**4.1. Desired characteristics**

**4.2 Algorithm characteristics**

The meta-review in Table 1 compares common supervised learning algorithms across characteristics of the two domains mentioned in the previous section: data structure and application problem. This review is cross-referenced with the relevant characteristics of the current application problem (startup funding prediction) and data set (CrunchBase) to evaluate and rank the candidate learning algorithms. Each candidate learning algorithm, critique their advantages and disadvantages, and present evidence of their effectiveness in relevant applications.



**Table 1:** A meta-review of common supervised learning algorithms with respect to the constraints of application problems and assumptions made about the structure of the data sets. Dots are colour-coded with respect to performance (green: best, red: worst, yellow: in-between). Criteria are cross-referenced with respect to the current application problem and proposed data set (ticks: relevant criteria, crosses: irrelevant criteria). Performance is coded, totalled and ranked to produce a measure of appropriateness of fit to the current application problem and data set. NB: Naïve Bayes, LR: Logistic Regression, K-NN: K-Nearest Neighbours, DT: Decision Trees, RF: Random Forests, SVM: Support Vector Machines, ANN: Artificial Neural Networks.

**4.2.1 Naïve Bayes**

Naïve Bayes is a simple generative learning algorithm. It is a Bayesian Network that models features using a directed acylic graph, with the assumption that all features are independent. While this assumption is generally not true, it simplifies estimation. Naive Bayes converges quicker than discriminative models like Logistic Regression, requiring less training data. Naïve Bayes can be a good choice for small datasets with high dimensionality and sparse datasets as it estimates each feature independently. Even for larger datasets, Naïve Bayes has been found to sometimes outperform more complex machine learning algorithms because it is reasonably robust to violations of feature independence, at least with regards to classification (Niculescu-Mizil & Caruana, 2005). However, Naïve Bayes is known to be a poor estimator of class probabilities, especially with highly correlated features (Zhang, 2004). Naïve Bayes was used alongside Logistic Regression, Decision Trees and Support Vector Machines to predict success in equity crowdfunding campaigns on the AngelList data set (Beckwith, 2015). None of these models performed well (Mean F1-Score for positive classes = 0.26), especially for startups that ultimately received funding, though Naïve Bayes (AUC = 0.73) and Logistic Regression (AUC = 0.74) produced the best results. The author suggested insufficient features were captured in the training set, particularly features of the company’s product, its marketing content and quality, and the founders’ professional networks. The most commonly selected features in their models were funding history, located in San Francisco, and Twitter presence respectively.

**4.2 Logistic Regression**

Logistic Regression is a regression model. Logistic Regression fits probabilities for the response levels using a logistic function. The regression coefficients are usually estimated using maximum likelihood estimation.A strength of Logistic Regression is that it is trivial to adjust classification thresholds depending on the problem (e.g. in spam detection (Hastie et al., 2008), where it is important that specificity is high). 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. However, Logistic Regression tends to underperform against more complex algorithms like Random Forest, Support Vector Machines and Artificial Neural Nets in higher dimensions (Caruana et al., 2008). (EVIDENCE – CASE STUDY)

**4.3 K-Nearest Neighbours**

K-Nearest Neighbours is a common lazy learning algorithm. Lazy learning algorithms do not perform explicit generalisation, but compare new instances with instances from training stored in memory. K-Nearest Neighbours is based on the principle that the instances within a dataset will exist near other instances that have similar properties (Cover and Hart, 1967). 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 (Kotsiantis, 2007). 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 (Ahn & Kim, 2008). 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 managed to improve the K-Nearest Neighbours algorithm to the point where it outperformed the Artificial Neural Network.

**4.4 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 (Hunt et al., 1966) and Gini Index (Breiman et al., 1984). 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 (Caruana & Niculescu-Mizil, 2006). These shortcomings have been addressed with pruning mechanisms and ensemble methods like Random Forests, respectively. Decision Trees were compared with Naïve Bayes and Support Vector Machines to predict investor-startup funding pairs using CrunchBase social network data (Liang & Yuan, 2015). Decision Trees (CART implementation, Breiman et al., 1984) had the highest classification accuracy and the authors suggested they were particularly useful in this application because their reasoning could be easily communicated to startups.

**4.5 Random Forests**

Random Forests are an ensemble learning technique that constructs multiple unpruned decision trees from bootstrapped samples of the training data, using random feature selection. Prediction is made by aggregating the predictions of the ensemble. Random Forests exhibit a performance improvement over a single Decision Tree classifier and are among the most accurate learning algorithms (Caruana, 2006). In contrast to a single Decision Tree, Random Forests are less likely to overfit because of bootstrapping at the training data-level and at the feature-level. However, Random Forests are more complex than Decision Trees by nature, taking longer to create predictions and producing less interpretable output. Random Forests were used to predict private company exits using quantitative data from ThomsonOne (Bhat & Zaelit, 2011). 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 (Breiman, 2001). The authors suggested that meta-classifiers such as boosting and bagging also performed well and deserve investigation in future work.

**4.6 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 typically high accuracy (Caruana & Niculescu-Mizil, 2006), theoretical guarantees on limiting overfitting, and with an appropriate kernel they can work well even if data isn’t linearly separable in the base feature space (though this is an issue with a linear kernel). Support Vector Machines are computationally intensive and relatively 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 (Shin et al., 2005). Support Vector Machines were found to outperform Artificial Neural Networks at this task, especially because it was on a relatively small data set.

**4.7 Artificial Neural Networks**

Artificial Neural Networks are a computational approach based on a network of neural units (neurons) that loosely models the way that the brain solves problems. Neurons are usually classified into input units, which receive information to be processed; output units, where the results of the processing are found; and units in between known as hidden units which allow for high-dimensional approximation. There are several algorithms with which a network can be trained (Neocleous & Schizas, 2002). The most well-known and widely used learning algorithm to estimate the values of the weights is the back-propagation algorithm. Artificial Neural networks tend to be highly accurate but are very 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. (EVIDENCE – CASE STUDY)

**4.3 Evaluation of algorithms**

I evaluated common supervised learning algorithms for their suitability to the current application problem and data set, startup funding prediction using the CrunchBase dataset. There is no clear best-in-class algorithm. I expect Random Forests, Support Vector Machines and Artificial Neural Nets to produce the highest classification accuracies. An ensemble of these high-performing methods may also provide an accuracy improvement, though at the cost of computational speed and interpretability. Random Forests could be expected to slightly 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 might be preferable for early, exploratory analysis of the dataset. Ultimately, however, larger training sets and good feature design tends to outweigh algorithm selection.

**5. Conclusion**

**References**

**Bibliography**

**Xiang 2012**

In addition to the financial and managerial view of this problem, data mining and machine learning strategies were also explored (**Meador, Church, and Rayburn 1996; Ragothaman and Ramakrishna 2002; Slowinski, Zopounidis, and Dimitras 1997; Wei, Jiang, and Yang 2009**). Based on the Naive Bayes classification model, Wei et al. proposed a set of features (**Wei, Jiang, and Yang 2009**) that model a company’s technological quantities. They proposed to utilize the ensemble learning algorithm on resampled data to solve the problem of data skewness, resulting in a TP of 46:43% on 2; 394 companies out of which 61 actually got acquired. In another work, an expert system, ACQTARGET, was designed as a useful evaluation tool to classify firms into acquisition and non-acquisition target categories (**Ragothaman and Ramakrishna 2002**).

* **Meador, Church, and Rayburn 1996; - LR**
* **Ragothaman, Naik, Ramakrishnan 2003; - ES, MDA, LR/PR**
* **Slowinski, Zopounidis, and Dimitras 1997; -**
* **Wei, Jiang, and Yang 2009 – BN/NB**

Lastly, researchers have also studied business failures and bankruptcies. Among them, the first dated back to (Altman 1968; Beaver 1966) 1960s, which used empirical methods and proposed several financial ratios as features, giving rise to multivariate statistical analysis (Karels and Prakash 1987) and discriminant analysis (Deakin 1972) for this task. Since early 1990s, machine learning and data mining techniques dominated the domain of bankruptcy prediction, yielding a few representative works such as (**Olson, Delen, and Meng 2012; Cho, Hong, and Ha 2010; shik Shin, Lee, and jung Kim 2005; Wilson and Sharda 1994**). In (Wilson and Sharda 1994), Wilson and Sharda compared the predictive power of neural networks and discriminant analysis based on five financial ratios, and concluded that neural networks clearly outperformed the traditional discriminant analysis. Moreover, Shin et al. showed in (shik Shin, Lee, and jung Kim 2005) that the Support Vector Machine (SVM) also achieved a competitive performance especially with small training sets.

* **Olson, Delen, and Meng 2012;**
* **Cho, Hong, and Ha 2010;**
* **shik Shin, Lee, and jung Kim 2005;**
* **Wilson and Sharda 1994**

We also evaluated our approach using the Support Vector Machines (SVM) and Logistic Regression (LR). Limited by space, we refrain from reporting all the statistics. However, the finding here is that both SVM and LR were significantly outperformed by BN in terms of TP, with SVM using the radial basis function as the kernel. Although their FP is better than that of BN, BN still comes out the winner with respect to the overall performance. For example, the TP and FP on companies in the aggregate “computer” category for SVM and LR without topic features are 39:6%/0:1% and 2:8%/0:3% respectively, while BN achieved 59:9%/2:2%. This observation is not surprising due to the correlation among our features and the absence of a linear separator in the feature space that SVM and LR typically learn.

For M&A prediction, the most important aspect of an algorithm is its TP. We believe companies are very careful in acquiring others, using techniques like ours in this paper as guideline only. Therefore, no liability issues are involved as in other domains such as spam detection, and thus a relatively higher FP does not matter as much.

**Wei, Jiang, and Yang 2009**

As for the analysis methods applied to learn an M&A prediction model, logistic regression is the most common one [1], [6], [10], [13], [15]. Discriminant analysis and rule induction can also be found in some prior studies [15].

* **[1] Ali-Yrkko, Hyytinen, Pajarinen (2005)**
* **[6] Gugler, Konrad (2002)**
* **[10] Meador, Church, Rayburn (1996)**
* **[13] Pasiouras, Gaganis (0207)**
* **[15] Ragothaman, Naik, Ramakrishnan (2003)**

As mentioned, we formulate M&A prediction as a classification problem with two possible decision categories: M&A and non-M&A. To learn an M&A prediction model, we need to prepare a set of training examples of the two categories. Because it is easier to obtain non-M&A cases than M&A cases, a set of training examples tends to be highly asymmetric or skewed in the two categories. To deal with such skewness in a training set, we develop an ensemble approach for learning. Given a training set, the basic idea of our ensemble learning algorithm is that we sample k training subsets from the training set for training k base classifiers. Assume that we have n M&A cases and m non-M&A cases (where n < m) in the training set. Each training subset will comprise all of the n M&A cases and α×n randomly sampled non-M&A cases, where α is used to adjust the ratio of non-M&A and M&A cases in a training subset. Subsequently, an induction learning algorithm (specifically, Naïve Bayes classifier [11] in this study) is applied to build a base classifier for each training subset.

To predict a new (unseen) case (i.e., a pair of a bidder company and a candidate target company), each base classifier will produce the probability that the case belongs to the M&A category. We can employ the average method by averaging the probability predicted by each base classier to arrive at the overall probability. Consequently, if the overall probability is equal to or greater than 0.5, the case will be predicted as in the M&A category; otherwise, the non-M&A category. Alternatively, we can adopt the weighted average method to obtain an overall probability, where the weight of a base classifier depends on its prediction accuracy on the training set.

**Meador, Church, Rayborn 1996**

Logit analysis was expected to prove more powerful than multiple discriminant analysis (MDA) which was used in earlier studies by Stevens [13], Simkowitz and Monroe [12], and Wansley [14]. Discriminant analysis requires the data to have multivariate normal distribution and the dispersion matrices of the groups to be equal. Neter and Wasserman [7, p. 329] state “both theoretical and empirical considerations suggest that when the dependent variable is binary, the underlying relationship is frequently curvilinear.” In logit analysis, no assumptions need be made about the prior probability that the firm belongs to a specific group, and the assumptions of normal distribution and the equality of variances and covariances across groups are less critical.