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

W.M.R. Shelton

30 August 2016

Component: Research Proposal

Supervisors: Professor Melinda Hodkiewicz, Dr Tim French

Degree: BPhil(Hons) (24 point project)

University: The University of Western Australia

Background

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

One explanation for why startups are staying private for longer is the accelerating nature of global business. Startups, particularly those backed by VC firms, are expected to scale fast and require frequent rounds of fundraising coupled with centralized, quick decision making. Such flexibility is not afforded to public companies, due to strict reporting and compliance requirements [2].

Why does this waiting game matter? Principally, because it shifts value creation to the private markets. To put things in perspective, Microsofts market cap grew 500-fold following its IPO, but for Facebook to do the same now its valuation would have to exceed the global equity market. VC funding for late-stage startups is approaching all-time highs as investors are entering the private markets to seek higher returns [1].

Merger and Acquisitions (M&A) have far surpassed IPOs as the most common liquidity event for startup founders and investors. In 2015, five times as many US-based VC-backed startups were acquired compared to those that went public through an IPO [1]. Accordingly, startup founders and investors may be interested in predicting which startups are likely to be acquired and by whom. However,

M&A prediction is a challenging task in the private markets where there is a lack of publicly-available information.

M&A prediction techniques have been proposed in the literature but often share a few weaknesses for the application of startups in the private markets. Previous work has relied on relatively small data sets [3] because publicly-available information on private companies is scarce. In addition, previous work has focused on the financial or managerial features of potential targets [4] with little work on textual or social network features.

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

There are limitations to Xiang and colleagues' study: the CrunchBase corpus they studied was sparse and relatively small in 2012, only a few common binary classification techniques were tested, and their approach didn't consider IPOs or bankruptcies as potential outcomes. In addition, it is unclear how robust their classifiers are through time. The study could be extended by applying the topic modelling approach to other text corpora such as patent filings, or by attempting a social network link prediction model.

Aim

We aim to produce a supervised learning model that will accurately predict the acquisition of startups in the private markets. We will build on the study by Xiang and colleagues (2012) [5], introducing new features and classification techniques. In the previous study, True Positive rate (TP), False Positive rate (FP) and Area under the ROC curve (AUC) were the main evaluation metrics used (collectively, known as "accuracy").

Hypothesis 1 (H1) Xiang and colleagues (2012) [5] results can be replicated

H2 Introducing new classification techniques improves accuracy

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

H3 Introducing additional CrunchBase features improves accuracy

Xiang and colleagues' study used a total of 22 factual features from CrunchBase profiles. No feature selection process was documented. A recent similar study on AngelList (which has a sharing agreement with CrunchBase) used 85 features of which 11 were selected [6]. Of those 11 features, many were not included in Xiang and colleagues' model. It is plausible that broadening the feature space may result in an improved model.

H4 Introducing additional labels improves accuracy

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

H5 Using more recent CrunchBase corpora improves accuracy

Xiang and colleagues' study used a CrunchBase corpus from 2012. They found the corpus relatively sparse at the time. Since 2012, the CrunchBase corpus has significantly grown. The CrunchBase Venture Program has encouraged more investors to provide data, and the AngelList - CrunchBase data sharing agreement has helped too. It is plausible that a more recent CrunchBase corpus can provide a better basis for an accurate model.

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

Method

1. Replicate study by Xiang et al. (2012) [5]

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

   - Features:
     - Factual Features (CrunchBase)
       - Basic Features e.g. office location, company age
       - Financial Features e.g. investment per funding round
       - Managerial Features e.g. number of acquired companies by founders
     - Topic Features (TechCrunch articles)
   - Outcome: Acquired? (CrunchBase)
   - Processing:

    - Topic model - Latent Dirichlet Allocation (LDA)
    - Classification techniques
        * Bayesian Network (BN)
        * Support Vector Machines (SVM)
        * Logistic Regression (LR)

2. Test additional classification techniques
    - CART Decision Tree (CART) as in [6]
    - Restricted Bolzmann Machine (RBM) as in [6]
    - Random Forest (RF)
    - And other classification techniques
3. Expand the factual features set
    - Founder education (CrunchBase, Dec-2013) as in [6]
    - Founder employment (CrunchBase, Dec-2013) as in [6]
    - Founding team (CrunchBase, Dec-2013) as in [7]
    - And other factual features in the CrunchBase corpus
4. Incorporate other potential startup outcomes
    - Outcomes: Bankrupt, Acquired, Public
    - Classification techniques: One vs. all (OVA), All vs. all (AVA)
5. Test classifier robustness over different datasets
    - Original dataset from Xiang et al. (2012) [5]
    - CrunchBase readily-available snapshot (December 2013)
    - CrunchBase recent crawl (September 2016)
6. Extend topic modelling and introduce network features (stretch goal)
    - Domain-Constricted LDA model (TechCrunch articles) as in [8]
    - Patent similarity (Google Patents) as in [9]
    - Social network link prediction (CrunchBase) as in [10, 11]
    - And other types of features as time permits

Timeline

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

S:W	Date	Task
2:03	Fri 19 August	Draft proposal due
2:05	29 Aug - 02 Sep	Proposal defence to research group
2:07	Fri 09 September	CrunchBase corpora collected
2:SB	Fri 30 September	Draft literature review due
2:12	Fri 28 October	Revised proposal due
2:12	Fri 28 October	Literature review due
2:15	Fri 18 November	Replicated Xiang et al. study
1:01	Mon 27 February	Completed experiments (exc. stretch goals)
1:08	Fri 28 April	Draft dissertation due
1:10	Fri 12 May	Seminar title and abstract due
1:13	Mon 29 May	Final dissertation due
1:13	Fri 02 June	Poster due
1:13	29 May - 02 June	Seminar
1:17	Mon 26 June	Corrected dissertation due

Table 1: Proposed timeline

Software and Hardware Requirements

This project will be developed primarily in Python using scikit-learn, a free open-source machine learning library [12]. MySQL may be used to prepare datasets for processing. The system will be hosted on a public compute cloud, likely Amazon Web Services. A free academic license for CrunchBase has been requested.

References

[1] *National Venture Capital Association (NVCA) Yearbook*. Thompson Reuters, 2016.

[2] Wies, Simone, and Christine Moorman. "Going public: how stock market listing changes firm innovation behavior." *Journal of Marketing Research*. 2015

[3] Wei, Chih-Ping, Yu-Syun Jiang, and Chin-Sheng Yang. "Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach." *Workshop on E-Business*. Springer Berlin Heidelberg, 2008.

[4] Hongjiu, Liu, Chen Huimin, and Hu Yanrong. "Financial characteristics and prediction on targets of M&A based on SOM-Hopfield neural network." *2007 IEEE International Conference on Industrial Engineering and Engineering Management*. IEEE, 2007.

[5] Xiang, Guang, et al. "A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch." *ICWSM*. 2012.

[6] Beckwith, John Jack. "Predicting Success in Equity Crowdfunding.". 2016.

[7] Spiegel, Olav, et al. "Going it all alone in web entrepreneurship? A comparison of single founders vs. co-founders." *Proceedings of the 2013 annual conference on Computers and people research*. ACM, 2013.

[8] H. Yuan, et al., "The determinants of crowdfunding success: A semantic text analytics approach", *Decision Support Systems*. 2016.

[9] Huang, Lu, et al. "Identifying target for technology mergers and acquisitions using patent information and semantic analysis." *2015 Portland International Conference on Management of Engineering and Technology (PICMET)*. IEEE, 2015.

[10] Shi, Zhan, Gene Moo Lee, and Andrew B. Whinston. "Towards a better measure of business proximity: topic modeling for analyzing M&As." *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 2014.

[11] Yuxian, Eugene Liang, and Soe-Tsyr Daphne Yuan. "Investors Are Social Animals: Predicting Investor Behaviour using Social Network Features via Supervised Learning Approach.". 2013.

[12] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research*. 2011.