CHAPTER 1

Discussion

In previous chapters, we described how we designed and evaluated a novel data mining system for automated Venture Capital (VC) investment screening. In this chapter, we discuss the merits of this project with respect to our system's design, our system's performance, and its contributions to theory on VC investment.

1. System Design. We developed a data-mining system that provides automated VC investment screening. Our system leverages a comprehensive feature set from CrunchBase and PatentsView. We developed a pipeline optimisation system to address limitations in these data sources and adapt to their changes over time. Our system is semi-autonomous and designed so that the VC industry can adopt it with minimal further development.

2. System Performance. We evaluated the performance of our VC investment screening system. We found that our system's performance to be robust over time and better or comparable to previous results from the literature. However, we identified that our system does not adequately capture nuances of investment success. We outline ways to improve our system's performance in the future.

3. Model Evaluation. We developed a novel conceptual framework for VC investment decisions which guided our data source selection and feature creation. In evaluating our VC investment screening system, we also evaluated our conceptual framework. In that sense, we contribute a comprehensive, empirical study of startup performance. We discuss ways to build on our conceptual framework in future research.

1.1 System Design

The design of our automated Venture Capital (VC) investment screening system provides contributions to the literature including the use of data sources

CrunchBase and PatentsView, and the development of an adaptive pipeline optimisation system. In the following section, we discuss these contributions and areas for future development.

1.1.1 Data Collection & Preparation

Our system collects data from CrunchBase and PatentsView. Neither data source has been studied comprehensively in the literature. CrunchBase gained critical mass in 2012 and PatentsView was formed in 2015. CrunchBase and PatentsView offer a significant improvement regarding size and variety of features over previous data sources used in entrepreneurship research (e.g. surveys, interviews, closed datasets, etc.). However, we did find that collecting data from CrunchBase and PatentsView raised some issues that we needed to overcome.

We found the CrunchBase database to be highly sparse with many long-tailed features and irrelevant companies. These qualities have been identified in the literature previously [1] and are likely attributable to the crowd-sourced nature of CrunchBase. To address CrunchBase's dataset issues, we developed a classification pipeline with multiple pre-processing steps (as described in the next section). In addition, there is room for improvement in the implementation of our data collection system from CrunchBase. The current system downloads CSV dumps exported from the CrunchBase database. While this is an improvement on the status quo (manual data collection), a further improvement would be the development of a real-time connector to CrunchBase's API. CrunchBase publishes a daily changelog which our connector could use to request only information from companies that had changed. An entirely time-stamped database produced in this manner would also allow for greater analysis of temporal trends, and avoid the biases of our database slicing technique.

The primary issue we encountered during collection of PatentsView data was matching companies between PatentsView and CrunchBase. Companies often use variations on their names on these sources. We attempted to resolve this issue by standardising the names (e.g. removing suffixes and punctuation) and using Levenshtein distances to determine the likelihood that we were accurately matching the companies. We did observe a positive correlation between the number of patents filed by companies and their developmental stages which may suggest the matching was successful. However, PatentsView features did not contribute significantly to the models generated by our system despite previous studies that show patent filings to be key predictors of startup performance [2, 3]. Another issue that we encountered with PatentsView was that their API, unlike CrunchBase, does not provide a change-log. This limitation means our system

has to run a full database sweep (approximately 10 hours) to check its patent filing records are up-to-date. We hope PatentsView will add this functionality soon.

A key contribution of this project is our ability to capture the trajectory of startups and reliability of VC investment predictions through time. This contribution is only possible because of our database slicing technique, but this technique also raises concerns. We evaluated the slicing system leveraging both 'last-updated' and 'created-at' timestamps: using 'last-updated' timestamps excluded many recently-updated companies, whereas using created-at timestamps retained more features than is historically accurate. We decided to use created-at timestamps. While the relational database structure mediates this effect (e.g. acquisition and investment records have separate timestamps), this technique may inflate our system's performance on companies that have had many recent edits to their records. We attempted to evaluate this impact by comparing a CrunchBase database collected in December 2013 with a slice from our primary database collected in September 2016. We found minimal variance in the number of records found in each relation. However, because the database schema changed between 2013 and 2016, we were unable to determine whether the completeness of each respective record is similar. As we collect more database dumps, we will be able to evaluate this technique better.

1.1.2 Pipeline Optimisation

Previous studies that have applied data mining techniques to startup performance and VC investment have typically evaluated a few specific classification algorithms [4, 5, 6, 7]. In a novel contribution, we presented an adaptive pipeline optimisation process that provides greater accuracy and re-calibration of the system as the data sources change over time.

Pipeline creation performs a broad search and evaluation of candidate pipelines with varying hyper-parameters. This search is performed across the pre-processing steps of the pipeline and also the classification algorithms. We found classification algorithm tuning had the greatest impact on performance during optimisation. It appears little optimisation of the pre-processing steps were needed. In aggregate, the performance of the classification pipeline was not improved by the pre-processing steps. Interactions between the pre-processing steps and the classification algorithms (e.g. Random Forests are more resilient to low orthogonality than Naive Bayes) may have reduced this aggregate effect. Nonetheless, optimisation of the pre-processing steps should still improve the overall robustness of the optimisation process as the data set and prediction tasks change.

Our literature review suggested Random Forests would be the most successful classifier, followed by Artificial Neural Networks and Support Vector Machines. We found Random Forests and Logistic Regressions performed best and Artificial Neural Networks and Support Vector Machines underperformed. Random Forests may have outperformed the other algorithms due to its robustness to missing values and irrelevant features. Learning curves also revealed that, unlike most of the other classifiers, Random Forests was least likely to converge early, which suggests with larger training sets it should perform better. Logistic Regression. It was surprising that Support Vector Machines and Artificial Neural Networks underperformed the other algorithms. However, these algorithms are harder to tune accurately, so it may reflect that our search process was too limited. In production, we could perform the search process over longer iterations which would likely result in better performance from these algorithms.

The second step in our system's pipeline optimisation process is pipeline selection. In this component, we rank the candidate pipelines generated previously and evaluate the best pipelines over different dataset slices. This process ensures our final pipeline is robust in its performance over time. We do not observe significant variance in the pipelines on aggregate against the dataset slices, but there is variance in each pipeline's performance against the dataset slices. The former result suggests our pipelines produce models that are robust over time (which we reinforce in our later experimentation). The latter result justifies this step in our process. In our preliminary evaluation of this test, we selected the top ten candidate pipelines. Although there is still a strong positive correlation between the pipeline's initial ranking and their scores, we can see there are some individual deviations. Importantly, the top-ranked pipeline from the first stage has a lower median score than the second-ranked pipeline. These results suggest it is optimal to evaluate the top 3-5 candidate pipelines in this manner.

1.1.3 Automation & Efficiency

A key benefit of our proposed VC investment screening system is that it will reduce the amount of manual effort required before an investment decision. The implementation of the system described in this paper goes some ways to address this. Currently the system is semi-autonomous: it has little requirement for external input besides configuration of investment criteria (e.g. forecast window, developmental stage, etc.), but still runs on-demand, rather than continuously.

An improved implementation of the system would run in the background continuously, scheduling components of the system to run as needed to ensure the results are always optimised. The system currently takes a total of 46 hours to

complete. Most of the duration of the system is because of pipeline creation, which performs a large search across potential pipeline hyper-parameters. However, when placed into production, this component could infrequently be run (e.g. once per year) to ensure the pipelines remain optimally suited for the dataset. The next component of the system, selecting the most robust pipeline, could occur more frequently (e.g. once every month). The final component of the system, making predictions, could be evaluated whenever the system collects new data (e.g. once per day) because it only takes an hour to complete.

When we evaluated the learning curves of the pipeline selected by our system, we found our system's performance had converged for some target outcomes but for others (e.g. IPO, Acquisitions) it had yet to converge. This finding suggests our system could still benefit from a larger dataset. However, these results could also arise because our experimental configuration uses a pipeline that was optimised for predicting our base target outcome. We might find other classification pipelines yield better performance if the entire system runs for different target outcomes. A key advantage of our system is that, as we collect more data over time, our pipeline optimisation process will adapt to the nature of the dataset and select classifiers with less bias and more variance so that we may see Support Vector Machines or Artificial Neural Networks adopted by the system in the future.

1.2 System Performance

We evaluated the performance of our Venture Capital (VC) investment screening system across a range of datasets with different training dates, forecast windows, developmental stages, and target outcomes. In this section, we discuss the performance of our system across these domains, with a comparison to previous systems from literature. We also suggest some limitations of our experimental design.

1.2.1 Historical Datasets

Robustness over time is a critical attribute for our system, so VC firms can rely that models trained on historical datasets will be accurate into the future. We evaluated the robustness of our system's performance by training the system on datasets of different dates from 2012-2014. We evaluated the performance of the system using a variety of evaluation metrics. Across all evaluation metrics, we observed little variance in performance. We observed that performance variance

decreased as forecast windows became longer in duration. This trend may be because longer-term models use fundamental features that are less likely to vary over time. This trend may also be because there is greater variance in the dates of datasets with shorter forecast windows because our database slicing technique is restricted in how far back we can reverse-engineer data slices before there are not enough observations left to proceed.

1.2.2 Forecast Window

While predicting whether a startup is likely to exit in the future is useful, predicting the timing of that exit is also critical because the time taken to exit is inversely related to the rate of return that the investor receives. We evaluated our system's performance against forecast windows of 2–4 years and a variety of evaluation metrics. We observe a positive relationship between performance and length of forecast window. This trend suggests that it is harder to predict when a startup company will raise a funding round or exit than whether it will do it at all. Non-performance related factors (e.g. finding an investor with the right fit, requiring extra funding to enter a new market) may influence the timing of each activity. In the future, it would be interesting to explore forecast windows closer in duration to a typical VC investment horizon (3–8 years) [8]. As we discovered in our preliminary evaluation, the majority of companies that are going to raise funds at all will raise funds over a three-year period. However, exits seem to operate on a longer lead-time. It is unlikely that our ability to predict additional funding rounds would change over a longer forecast window but we would expect to see our performance at predicting exits continue to increase. We also observe that a relationship between the magnitude of performance improvement over longer forecast windows and how sensitive the evaluation metric is to both the imbalanced nature of the dataset and our bias towards the positive class. Accordingly, F1 Scores show the greatest performance improvement and Area under the Receiver Operating Characteristic (ROC) curve shows little variance. This finding justifies our decision to select F1 Scores as our primary evaluation metric.

1.2.3 Developmental Stage

VC investment decision-making is difficult. In our preliminary evaluation of our dataset, we discovered that even firms at late developmental stages (e.g. Series B, C) who have received multiple rounds of screening, only have about a 30% chance of raising additional funding or exiting. We investigated how the performance

of our system varies across target companies of different developmental stages ranging from Pre-Seed through Series D+. We found a positive trend between later developmental stage and performance. Later-stage companies tend to have more available features in our dataset which may explain this trend. Beckwith (2016) studied companies seeking equity crowd-funding (which maps to Pre-Seed in our classification system) and showed poor classification results for predicting whether a company would raise the equity crowd-funding round [4]. Beckwith's highest F1 Score was 0.33, which is in line with the results we gained for more difficult prediction tasks (e.g. predicting IPO over three years). Stone (2014) suggested that VC investment screening was simply not viable before Series A stage [9] whereas we observed performance up to an F1 Score of 0.58 for predicting funding rounds and exits by Seed stage companies. Bhat (2011) studied companies that had previously raised three VC rounds (Series C in our classification) and received comparable results to our system [5].

A discrepancy in the positive trend between developmental stage and performance is a slight decrease in our system's performance at Series D+. This finding may be because the model is primarily predicting exits at this developmental stage, rather than additional funding rounds, and exits are harder to predict. To investigate this discrepancy further, we split the datasets into their developmental stages and fit the model onto each of these sub-datasets individually. Pre-Seed companies make up most of our original dataset, and we see the smallest improvement for this stage. However, for Series D+ we see a large improvement, which suggests the features that predict Series D+ performance vary from earlier stages. Overall, this stage-specific fit method results in a performance improvement, despite each model having significantly fewer observations on which to train. This finding suggests that the underlying factors that influence startup investment performance for each developmental stage are significantly different.

1.2.4 Target Outcome

Ultimately, our system seeks to identify startup investment opportunities that will return their invested funds many times over within an investment horizon of a VC's fund (typically 3–8 years) [8]. However, our dataset has little information about valuations at funding rounds or during acquisitions because valuation is considered sensitive and confidential. Instead, we developed broader target outcomes as rough proxies for the underlying success of the investment. These outcomes include raising additional funds, being acquired, having an IPO and combinations thereof. We evaluated each of these outcomes separately to determine their effect on our system's performance. Our system is better at predicting more common events, so performs best at predicting additional funding rounds

and worst at predicting IPOs. The system's poor performance on IPOs may also be due to non-performance related factors that affect IPO timing, like financial market conditions. Surprisingly, our system is better at predicting whether a company will exit, than whether it will exit or gain additional funding in combination. Other factors interacting with target outcome may be causing this effect. For example, most companies that exit are in their later stages of development, while most companies that raised additional funding rounds are in earlier stages of their development. Our system may not be capable of learning to disentangle these factors when predicting a combined outcome.

While our target outcomes provide a proxy for investment success, some nuances are not captured by these outcomes. While most funding rounds are generally at higher valuations than the previous round, some funding rounds are not — these are termed 'down-rounds'. Likewise, although most acquisitions are at higher valuations, sometimes they are not, and are instead used to recruit many of the staff that worked at the startup — these are sometimes termed 'acqui-hires'. As our publicly-sourced dataset has little information about valuations at funding rounds or during acquisitions, our system has little ability to distinguish between successful activity and down-rounds or acqui-hires. These discrepancies limit the performance of our system. Appendix **??** presents four case studies that highlight the nuances of our system's performance. In future, the application of sentiment analysis to media coverage of funding rounds, acquisitions or IPOs may indicate whether the raise or exit was genuinely successful.

1.2.5 Experimental Design

Our experimental design involved evaluating the performance of the system across a range of variables, including the size of each training set, date of each training set, duration of forecast window, company developmental stage, and target outcome. For each of these experiments, we manipulated these variables during the model fit and prediction step of our system design. However, to reduce the time taken by our experiments, we used the same optimised pipeline for each experiment (for the configuration, see Appendix **??**). This pipeline optimisation step takes the vast majority of time of our system (84.8%). By using a pipeline optimised for different objectives we are likely to have under-reported the performance of our system. In future research, it would be interesting to determine the extent to which the results of our pipeline optimisation changes with respect to these variables and the extent to which our results improve.

1.3 Model Evaluation

As a by-product of the evaluation of our system, we were also able to provide a comprehensive study of the determinants of startup investment performance. From our literature review, we developed a conceptual framework for startup investment performance, based on previous work by Ahlers and colleagues [10]. Our conceptual framework proposed that Venture Capital (VC) investment decisions have two primary components: startup potential and investment confidence. We decomposed these components into 15 factors identified in previous empirical studies in the literature. Our system evaluates many of these factors. Through our experimentation on the system, our system generated models that describe features associated with startup investment performance over time, with different forecast windows, developmental stages, and target outcomes.

1.3.1 Historical Datasets & Forecast Window

Within the VC industry, there is a commonly-held belief that startup performance is too volatile and qualitative to be predicted with any real accuracy using data mining techniques [9]. The models generated by our system did not provide evidence to support this assertion. We evaluated our system by training it on a range of historical datasets from 2012-16 across varying forecast windows. We found that models generated by the system were robust to changes in training date, with a standard deviation of less than 1% of the total normalised feature weights. We found a similar trend with respect to forecast window, with little variance for periods of 2–4 years. Despite this, we previously observed a positive relationship between system performance and forecast window length. Together, these findings suggest that for a given set of independent variables (developmental stage and target outcome), our models of startup investment performance are stable over time. This finding is in contradiction to the widely held within the VC industry that the factors that influence startup investment performance change over time. Admittedly, our models do not perfectly predict startup investment success, and dynamic factors not incorporated in our system may cause this margin of error. If not to a decision-making degree of performance, our results still suggest features that correlate with startup investment success are predictable and stable.

1.3.2 Developmental Stage

Previous studies have largely neglected to investigate how models of startup performance change across the startup development life-cycle. Most research in this field has studied either early-stage companies [4, 9, 11, 10], or much later-stage companies [5]. We believe that this is the first comprehensive study that takes a broader approach. We found considerable variation in models developed for companies of different stages. We find that *Advisors* is a more important factor to earlier stage companies than late stage companies. One explanation for this finding that successful startups recruit experienced influential advisors earlier to fill gaps in their experience but by the time these startups reach later rounds, this advantage is lessened by the influence of investors. We found that *Investors* becomes more important as companies approach an exit (Series D+). This finding may be due to influential investors having more experience at going through exit processes, and leveraging their contacts to make the process easier. We found that *Executives*, and *Founders* factors are important in Pre-Seed companies, which has previously been substantiated in the literature [4, 12]. This finding may be because at early stages of a company's development there is little to rely on aside from the previous experience and skill-set of the founding team and staff. Finally, we found that the *Economy* factor was most important at the Pre-Seed stage and has little effect at other stages. We did not expect economic factors to have a large effect on startup performance because, in comparison to larger, more established companies, startups are flexible, agile and have more focused markets. One explanation for the greater impact of the *Economy* factor at the Pre-Seed stage is that perhaps more people leave established companies to launch startups when the economy is doing poorly, but these companies do not survive longer than the Pre-Seed stage.

1.3.3 Target Outcome

The startup landscape is constantly changing and so is what is considered a successful VC investment. Amazon went public in 1997, just two years after its Series A, at a market capitalisation of $440M. Contrast that with Uber, which remains private six years on and recently raised $3.5B at a $59B pre-money valuation. While previous studies separated funding raises [4, 11, 10, 12] from predicting exits [5], we decided to evaluate our system across a variety of target outcomes that reflects this changing landscape. There is considerable variance between the models generated by our system to predict additional funding rounds, acquisitions and IPOs. The greatest predictor of additional funding rounds was *Advisors*, followed by *Executives*. The greatest predictor of acquisition was *Ex-*

ecutives, followed by *Investors* and *Founders*. The greatest predictor of Initial Public Offering (IPO) was *Investors*, followed by *Executives* and *Funding*. There is substantial evidence to support the finding that advisors, founders and executives predict startup performance [3, 13, 4]. Although a distinct model predicts each target outcome, all models are weighted towards human capital features. The high proportion of early stage companies in our dataset may inflate this effect because we would expect early stage companies to rely more heavily on human capital as they have few other resources. Aside from human capital, *Investors*, which covers the reputation and track record of invested VC firms, was a good predictor of acquisitions and IPOs. Investors who have a strong track record probably have had more experience helping portfolio companies exit and have more connections to leverage in this process. Most previous studies of startup funding focus on early-stage companies whereas studies that focus on exits tend to investigate later-stage companies. Our study bridges this divide in a way that is more holistic and more useful for investors.

1.3.4 Limitations

1.3.4.1 Missing Features

While CrunchBase and PatentsView provide features that cover much of our conceptual framework, there are factors we were unable to evaluate in this implementation of our system. Missing factors include media coverage, strategic alliances and financial performance. While it would likely be a significant factor in our models, we do not expect to be able to source financial information for our dataset in the future. In fact, we consider it to be a key benefit of our system that it can perform accurately without detailed financial information. The paucity of available financial data is what makes VC investment screening distinct from other fields of finance. Collecting data to support the other missing factors may be easier to source in the future. CrunchBase API provides an archive of media coverage on each startup company, so connecting directly to the CrunchBase API (rather than the CSV-dumps) would give access to this feature. Social influence is harder to capture because historical records of social media activity are hard to find. CrunchBase tracks whether companies have social media profiles but does not provide time-stamps for this information so we cannot use the information to create historical records. There are some Twitter historical data services, but these are expensive to use. Finally, strategic alliance information (e.g. with suppliers or universities) is not a typical feature that is recorded but could be engineered through textual analysis on media coverage. We should investigate these features in future work.

1.3.4.2 Simple Features

While our features covered a broad conceptual framework, we derive many of features from simple models (e.g. a company's location, the age of a company, the amount of funding a company has raised). While these types of features are consistent with most other studies in this field, the simplicity of these features may have reduced our system's ability to represent more complex factors. There is preliminary research that features derived from more complex models like semantic text features (e.g. keyword analysis from patents) [2, 11] and social network features (e.g. networks of social influence) [14, 15, 16] are significant predictors of startup performance. The factors that could benefit most from these features (e.g. patents and social influence) were under-represented in the models generated by our system, which supports this line of reasoning. There are disadvantages, however, in adopting more complex, dynamically generated features. A key contribution of this project was our ability to test our system's robustness by training on different data sets and then comparing the results. This process would not be possible if we did not maintain a consistent feature vector — we would be unable to train and test on different datasets because the features would not align.

Finally, while our project did incorporate longitudinal analyses (with respect to evaluating our system's performance against historical datasets and for different forecast windows), we did not specifically analyse temporal patterns in the trajectory of startups. Temporal analyses are an interesting area for future research. For example, probabilistic networks (e.g. Markov networks) could be used to represent the sequence and timing of different startup activities (e.g. media coverage, funding rounds, IPOs, hiring, etc.). Machine learning techniques could be applied to these networks to learn patterns of activity that are likely to lead to investment success. This technique has the potential to provide finer predictions than our cross-sectional models. It was probably not viable to apply these types of techniques to private companies in the past because not enough data was available. However, this project shows that CrunchBase and PatentsView cover a considerable feature set, particularly for later stage startups, and so this area deserves further investigation.

Bibliography

[1] Xiang, G., et al. A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence. 2012.

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

[3] Baum, J. A., and Silverman, B. S. Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing* 19, 3 (2004), pp. 411–436.

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

[5] Bhat, H. S., and Zaelit, D. Predicting private company exits using qualitative data. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2011, pp. 399–410.

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

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

[8] Gompers, P. A. Optimal investment, monitoring, and the staging of venture capital. *The Journal of Finance* 50, 5 (1995), pp. 1461–1489.

[9] Stone, T. R. Computational analytics for venture finance. PhD thesis. University College London, 2014.

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

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

[12] An, J., Jung, W., and Kim, H.-W. A Green Flag over Mobile Industry Start-Ups: Human Capital and Past Investors as Investment Signals. In: *PACIS 2015 Proceedings*. AIS Electronic Library, 2015, p. 67.

[13] Gimmon, E., and Levie, J. Founder's human capital, external investment, and the survival of new high-technology ventures. *Research Policy* 39, 9 (2010), pp. 1214–1226.

[14] Werth, J. C., and Boeert, P. Co-investment networks of business angels and the performance of their start-up investments. *International Journal of Entrepreneurial Venturing* 5, 3 (2013), pp. 240–256.

[15] Yu, Y., and Perotti, V. Startup Tribes: Social Network Ties that Support Success in New Firms. In: *Proceedings of 21st Americas Conference on Information Systems*. 2015.

[16] Cheng, M., et al. Collection, exploration and analysis of crowdfunding social networks. In: *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*. ACM. 2016, pp. 25–30.