CHAPTER 1

Discussion

In previous chapters, we outlined 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 and performance, and its contribution to theory on startup investment performance.

1. System Design. We developed a data mining system that provides automated VC investment screening. Our system uses data collected from CrunchBase and PatentsView. We found that CrunchBase and PatentsView databases are large, comprehensive, and growing. However, CrunchBase records are sparse, long-tailed, and require cleaning. PatentsView features have a positive impact on the system's performance. Our system generates a classification pipeline optimised to the dataset. We found classification algorithm tuning had the greatest impact on performance during optimisation. In particular, Random Forests and Logistic Regressions were the most successful classifiers, Support Vector Machines and Artificial Neural Networks underperformed. We found that the top 3-5 pipelines generated from the optimisation process should be checked for their robustness over time. A limitation of our design is that the dataset slicing technique may introduce biases related to artificial completeness. Our system is semi-autonomous but could be made fully autonomous with further development.

2. System Performance. We evaluated the performance of our VC investment screening system. We found that our system's performance is robust with respect to historical datasets (for 2012-16), which makes it suitable for forward-looking predictions. We found that the performance of our system was better or comparable to previous results from the literature. Performance was positively related to longer forecast window (for a period of 2-4 years), later developmental stage (e.g Series C, Series D+), and breadth of target outcome (e.g. Exit). Our system's observed performance may have been improved further if we had performed pipeline optimisation separately

for each experiment. A limitation of our system is that some nuances of investment success are not fully captured (e.g. down-rounds, acqui-hires). However, we still believe our system has performance practical for use in investment screening.

3. Model Evaluation. We developed a conceptual framework for startup investment performance based on the literature. This framework guided our data source selection and feature creation. We evaluated the framework using our data mining system. We found that models of startup investment performance generated by our system are robust with respect to time (for 2012-16) and forecast window (over 2-4 years), and vary with respect to developmental stage (e.g. Seed, Series A) and target outcome (e.g. IPO, Acquisition). Our system does have some limitations. We were unable to represent all factors from our conceptual framework (e.g. financial information), we may have understated some factors by not using more complex features (e.g. temporal relationships), and we were unable to generate structural models. Despite these limitations, we believe this project has made contributions to models of startup investment performance.

1.1 System Design

We developed a data mining system that provides automated Venture Capital (VC) investment screening using data collected from large online databases, CrunchBase and PatentsView. This system provides a multi-stage pipeline optimisation process that can automatically adapt to changes in the dataset or prediction task over time.

1.1.1 Data Collection

Our system uses data collected from CrunchBase and PatentsView online databases. Both data sources are fairly new (CrunchBase gained critical mass in 2012 and PatentsView was formed in 2015) and neither have been studied frequently in the literature. These data sources offer a significant improvement in terms of size and variety of features over previous data sources used in this research field (e.g. surveys, interviews, closed datasets). However, we found CrunchBase was sparse, with many long-tailed features, and other abnormalities, and required considerable cleaning to remove irrelevant companies. We addressed these issues with a number of pre-processing steps in our classification pipeline. Unlike CrunchBase, PatentsView is a government-regulated data source, so it has fewer data quality

issues. The greatest issue we had with PatentsView was matching companies between PatentsView and CrunchBase because companies often use variations on their names or have sub-entities. However, PatentsView data still produced a broad performance improvement to our system.

Our system converts the data we collect into historical datasets by using timestamps. While this technique provides significant benefits to our system, it also raises some concerns. We performed preliminary testing of our dataset slicing technique using last-updated timestamps and realised this would remove too many recently-updated records from the dataset. Instead, we used created-at timestamps which retain more records, but at the cost of possibly using features not originally available. While the impact of this effect is mediated by the relational database structure (e.g. acquisition and investment records have separate timestamps), it may artificially inflate historical results for companies that have had many later edits to their records. We tried to evaluate the impact of this technique by comparing a CrunchBase database collected in December 2013 with a slice from our primary dataset collected in September 2016. We found there was only minimal variance in the number of records found in each relation. However, because the database schema changed dramatically between 2013 and 2016, we were unable to determine whether the completeness of each respective record is similar. As we collect more original database dumps, we will be able to better evaluate this technique.

The current implementation of our system downloads CSV-dumps of the CrunchBase data and connects to the PatentsView API. While both of these data collection methods are an improvement upon the status quo (manual data collection) they can be improved further. In the earlier stages of our project we developed a connector to CrunchBase's API that provided real-time access to their database. CrunchBase publishes a daily change-log which our connector could use to request only information from nodes that had changed. Although we abandoned this approach because of time constraints, it deserves further investigation. A fully time-stamped database produced in this manner would also allow for greater accuracy and analysis of temporal trends, and avoid the issues raised by our current dataset-slicing technique. Although our current method of collecting data from PatentsView is through their API, unlike CrunchBase PatentsView does not provide a change-log. This limitation means we have to run a full database sweep (approximately 10 hours) whenever we want to check our patent filing records are up-to-date. We hope PatentsView will add this functionality in the near-future.

1.1.2 Pipeline Optimisation

A key component of our system is the pipeline optimisation process. While previous studies in this field have applied a few specific classification algorithms, we developed a pipeline optimisation process with the aim of greater accuracy and re-calibration of the system as the dataset changes over time. Our pipeline optimisation process is divided into two steps: pipeline creation and pipeline selection.

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 very 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. However, it is likely the effect of the pre-processing steps was also dependent upon the classification algorithm (e.g. Random Forests are more resilient to low orthogonality than Naive Bayes), which may have reduced the aggregate effect. Nonetheless, optimisation of the pre-processing steps should still improve the overall robustness of the optimisation process as the dataset 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 far more difficult to accurately tune than these other algorithms, 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 (finalist pipelines) over a number of different dataset slices. This process ensures our final pipeline is robust in their performance with respect to time. We don't observe significant variance in the pipelines on aggregate against the dataset slices, but there is variance within the individ-

ual pipelines. The former result suggests our pipelines produce models that are robust with respect to time (which is reinforced in the final evaluation). 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 pipelines initial ranking and their scores, we can see there are some individual deviations. Importantly, the top-ranked pipeline from the first stage actually has a lower median score than the second-ranked pipeline. These results suggest 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 prior to the investment decision-making process. 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 this time is taken by the pipeline creation component, which performs a large search across potential pipeline hyper-parameters. However, when placed into production, this component could be run infrequently - perhaps once per year - to ensure the pipelines being used are still optimally suited for the dataset. The next component of the system, selecting the most robust pipeline, could occur more frequently - perhaps once every month - and the final component of the pipeline, making up-to-date predictions, could be evaluated every time new data is fed into the system (perhaps once per day) because it only takes an hour.

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 suggests our system could still benefit from a larger dataset. However, these results could also arise because our pipeline was optimised for predicting our base target outcome, and if the entire system was performed on the different target outcomes we might find other classification pipelines yield better performance. 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 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 comparison to previous systems from literature. We also discuss the limitations of our experimental design.

1.2.1 Time Slices

We evaluated the robustness of our system's performance by training the system on datasets of different dates from 2012-2014. Robustness with respect to 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 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 seems reasonable as we would expect longer-term models might be based on more fundamental features that are less likely to vary with respect to time. This trend may also be due to greater variance in the dates of datasets with longer forecast windows. This is because we are restricted in how far back in time we can reverse-engineer data slices without reducing the training size by too much.

1.2.2 Forecast Window

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. This seems like a reasonable proposition as 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 (5-8 years). If we decompose the trend further, we see that there is a relationship between the magnitude of the performance improvement

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 show little trend.

1.2.3 Developmental Stage

We compared performance of our system 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. This is possibly a product of later-stage companies having more complete feature vectors. Beckwith (2016) studied companies seeking equity crowd-funding (which maps to Pre-Seed in our classification system) and showed poor classification results even just predicting whether a company would raise the equity crowd-funding round, let alone exit at a later date [1]. Stone (2014) suggested that VC investment screening was simply not viable prior to Series A stage [6]. Bhat (2011) studied companies that had previously raised three VC rounds (Series C in our classification) and received comparable results to our system [2].

A discrepancy in the positive trend between developmental stage and performance is a slight decrease in our system's performance at Series D+. This decrease may be because the model is primarily predicting exits at this 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 broad performance improvement. This is despite each model having significantly less observations to train on, which suggests that the underlying factors that influence startup investment performance for each 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). 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. As we would expect, 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 actually better at predicting whether a company will exit, than whether it will exit or gain additional funding. This requires further research.

While our target outcomes provide a rough proxy for investment success, there are nuances that can't be 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 performed at higher valuations, sometimes they are not - these are often 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. In Appendix**??** we present four case studies that highlight the nuances of our system's performance. In future, applying sentiment analysis to media coverage of funding rounds, acquisitions or IPOs may indicate whether the activity was genuinely successful.

1.2.5 Experimental Design

Our experimental design involved evaluating the performance of the system across a range of variables, including size of training set, date of 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 are 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 [**ahlers2012**]. Our conceptual framework posited that startup investment decisions are based on two primary factors: startup potential and investment confidence. We decomposed these factors further into 15 features that were identified in previous empirical studies in the literature. Many of these features are evaluated by our system. 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 Time Slices

We evaluated our system by training it on a range of historical datasets from 2012-16. We found that models generated by the system were robust to these changes in training date, with a standard deviation of less than 1% of the total normalised feature weights. This suggests that for a given set of independent variables (forecast window, developmental stage and target outcome), models of startup investment performance are stable over time. This is somewhat in contradiction to the widely held within the Venture Capital (VC) industry that the factors that influence startup investment performance change over time. Admittedly, our models do not perfectly predict startup investment success and this margin of error might be caused by dynamic factors not incorporated in our system. 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 Forecast Window

Robust with respect to forecast window (over 2-4 years)

1.3.3 Developmental Stage

Variable with respect to developmental stage (e.g. Seed, Series A)

1.3.4 Target Outcome

Variable with respect to target outcome (e.g. IPO, Acquisition)

, probably because Acquisitions make up a large proportion of Exits in our database

1.3.5 Future Research

1.3.5.1 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, social media influence, 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, it is 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. Data to support the other missing factors may be easier to source in the future. The 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 media influence is more difficult, because historical records are hard to find. CrunchBase does track whether companies have social media profiles, but does not provide time-stamps for this information which limits our ability to create historical records. There are a number of 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. These features should be investigated in future work that builds on our system.

1.3.5.2 Homogeneous Features

We incorporated features that covered a broad framework into our models, but the nature of these features was largely homogeneous. Most features were fairly basic: booleans, counts, summations, averages. This is in line with most of the prior work in this field that focused on basic company features (e.g. the headquarters' location, the age of the company) [1, 4]. However, there is preliminary research that has looked into using semantic text features (e.g. patents, media)

[5, 9] and social network features (e.g. co-investment networks) [7, 8, 3] with some success. We expect a model that includes semantic text and social network features alongside basic company features could lead to better startup investment prediction. However, a disadvantage of making models with dynamically generated features is that it becomes difficult to train and test on different datasets because the features do not align. This would have limited our ability to test the robustness of our system in the fashion that we did, training on different datasets and then comparing the results. Finally, aside from evaluating our system's performance against historical datasets and for different forecast windows, we did not incorporate temporal relationships into our system. However, in other forms of finance, like equity markets, time-series analysis is relied on heavily. Perhaps future research can look at the temporal relationships between different startup activities (e.g. media coverage, funding rounds, IPOs etc.) and how this chain of activity might predict investment success.

Bibliography

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

[2] Bhat, H. and Zaelit, D. "Predicting private company exits using qualitative data". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Huang, J., Cao, L., and Srivastava, J. Vol. 6634. Lecture Notes in Computer Science. Berlin: Springer, 2011, pp. 399–410.

[3] Cheng, M., Sriramulu, A., Muralidhar, S., Loo, B. T., Huang, L., and Loh, P.-L. "Collection, exploration and analysis of crowdfunding social networks". In: *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web*. ACM. 2016, pp. 25–30.

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

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

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

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

[8] 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.

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