CHAPTER 1

Evaluation

We believe it is possible to produce an improved Venture Capital (VC) investment screening system. In Chapter **??**, we described the design and development of such a system. Our system identifies startup companies likely to receive additional funding or exit in a given forecast window. In this chapter, we perform a series of experiments to evaluate our system against criteria of practicality, robustness and versatility.

1. Practicality. Our system is designed to require minimal user input. Therefore, by virtue of its design, the proposed system is more efficient than current systems. We also explored the time profile of our system. An indicative implementation of our system takes 46 hours to run.

2. Robustness. We observed minimal variance in the performance and types of models generated by our system across training sets of various dates. We also evaluated the learning curves of our system and identified that our system is likely to adapt and perform better as the data sources grow over time.

3. Versatility. We assesed our system's ability to screen investment candidates at different developmental stages, for different target outcomes and across different forecast windows. Our system's performance is positively related to longer forecast windows (for 2-4 years) and later developmental stage (e.g. Series B, Series C).

1.1 Experimental Design

In this chapter, we evaluate models generated by our system while varying a number of other factors. This evaluation process is depicted in Figure 1.1. The optimised pipeline is fit to a training dataset to generate a model. The model

is applied to a test feature vector to produce predictions. We scored these predictions against truth values derived from the held-out test database (collected in April 2017). This process is performed multiple times to evaluate the three primary criteria derived from our literature review: efficiency, robustness and predictive power. The configuration of the system during our experiments is detailed in Appendix **??**.

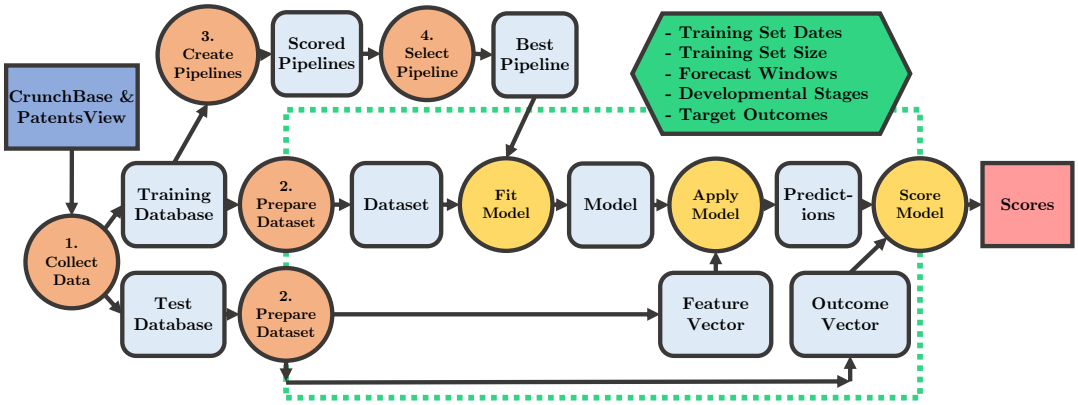

Figure 1.1: Pipeline evaluation overview. Training and test datasets are generated according to the experimental configuration: varying with respect to training set dates, training set size, forecast window, developmental stage, and target outcomes. Legend: dark blue square = input, orange circle = system component, yellow circle = process, light blue rounded square = intermediate, red square = output, green hexagon: iterative process / search space.

1.1.1 Baseline Analysis

Before we evaluated our system, we performed preliminary analyses to determine the baseline trends and distributions of company outcomes in our database.

First, we explored company outcomes by forecast window. We applied the same system of reverse-engineering time slices that we used in previous experiments on robustness, but this time we varied the time passed between our feature vector and outcome vector (i.e. the forecast window). We combined pair-wise datasets of each year from 2012-2016 and explored the proportion of companies that raised additional funding, were acquired or had an Initial Public Offering (IPO).

Figure 1.2 shows how company outcome varies with respect to the forecast window (time between the observed features and the measured outcome). We observe a positive relationship between length of forecast window and company

outcome. For additional funding rounds, this relationship disappears after three years. Few companies appear to have exited or raised funds over a period of less than two years so we will focus our experimentation on forecast windows of 2-4 years.

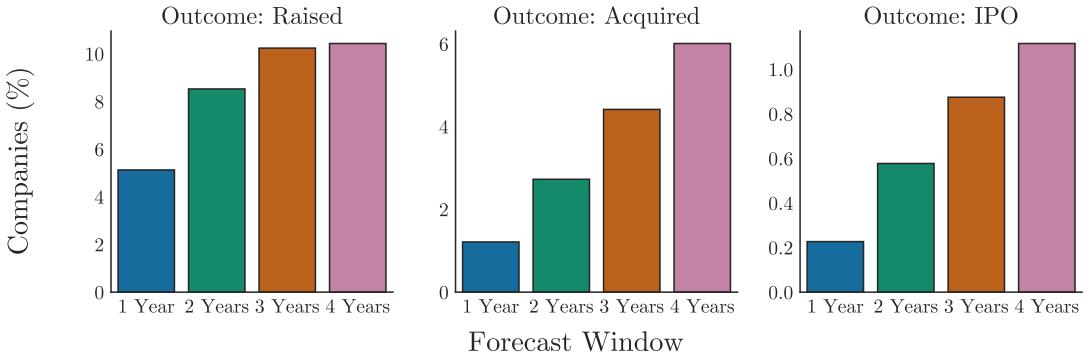

Figure 1.2: Outcomes by forecast window.

We also explored how company outcomes vary with respect to development stage, shown in Figure 1.3. We see a broad positive relationship between developmental stage and likelihood of further funding rounds and exits, which we would expect as at each stage there is higher market traction and scrutiny from investors. The variance between the outcomes of different developmental stages suggests that in our experimentation we should investigate how our system predicts each stage independently, as well as in aggregate.

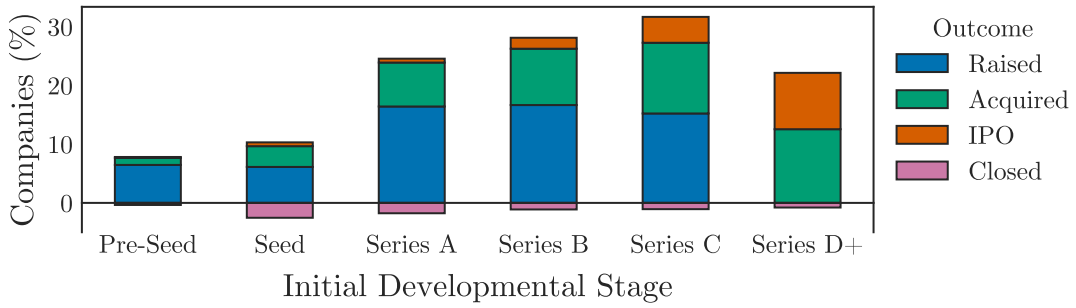

Figure 1.3: Outcomes by developmental stage.

1.1.2 Evaluation Metrics

While Area under the Precision-Recall (PR) Curve was used to guide the development of our system, in evaluation of our system's performance we primarily use

F1 Scores. An F1 score is the harmonic mean of recall and precision at points on the PR curve. In this sense, the Area Under Curve (AUC) measure provides an overall evaluation of a classification system, whereas the F1 Score evaluates a set of predictions. For investment screening, we are more sensitive to classification performance for the positive class (companies that have been successful in raising further funding or achieving an exit), so thereafter, when we refer to F1 Score, we refer to the F1 Score for this class alone. We also present Matthews Correlation Coefficient (MCC) in some of our analyses. MCC is a measure of the correlation between the observed and predicted binary classifications. The MCC should produce similar results to an F1 Score that incorporates performance across both classes.

1.2 Practicality

The Venture Capital (VC) industry requires more efficient forms of investment analysis, particularly in surfacing and screening. These processes are currently performed through referral, Google search, industry papers and manual search of startup databases. Our automated system is more efficient than these methods because it is designed to involve minimal user input. That aside, we assess the time profile of our system to determine whether it is practical for use in the VC industry.

Unlike other forms of finance, like equity or derivatives trading, VC operates on a much longer timeframe – deals close over weeks, rather than minutes. This has two key disadvantages: VC firms have higher management costs because they spend more time screening investments and startup founders waste precious time negotiating with investors when they could be building their businesses. Automated systems could decrease the time taken to generate investment opportunities.

An indicative time profile of the system is shown in Table 1.1. At the highest-level, this configuration of the program takes 46 hours to complete on a modern desktop PC. When we further break this time down by system component, the vast majority of time (84.8%) is taken up by the initial pipeline creation component. This time is due to the pipeline optimisation process – the model is fit and scored over 500 times on different classification algorithms and parameters. Scoring takes a long time because, in this case, it also involves generating learning curves for reporting, which is another cross-validated process.

Function	Cycle (s)	Cycles (N)	Time (s)	Time (m)	Time (h)
Generate Dataset (CV)	1,800	1	1,800	30	0.5
Prepare Feature Dataset	1,200	1	1,200	20	0.3
Prepare Outcome Dataset	180	1	180	3	0.1
Merge Datasets	360	1	360	6	0.1
Finalise Dataset	60	1	60	1	0.0
Fit and Score Model[1]	265	525	139,125	2,319	38.6
Fit Model	15	525	7,875	131	2.2
Score Model	250	525	131,250	2,188	36.5
Subtotal: Create Pipelines			140,925	2,349	39.1
Get Finalist Pipelines	5	1	5	0	0.0
Generate Dataset (CV)	1,800	5	1,800	30	0.5
Fit and Score Model[2]	265	75	19,875	331	5.5
Select Best Pipeline	5	1	5	0	0.0
Subtotal: Select Best Pipeline			21,685	361	6.0
Generate Dataset (Training)	1,800	1	1,800	30	0.5
Generate Dataset (Test)	1,800	1	1,800	30	0.5
Fit Model	30	1	30	1	0.0
Make Predictions	5	1	5	0	0.0
Subtotal: Fit and Make Predictions			3,635	61	1.0
Total			166,245	2,771	46.2

Table 1.1: System time profile. All times are indicative based on averages from system logs. Notes: [1] Cycles involve 25 search iterations, 3 cross-validated folds, and 7 classification algorithms. [2] Cycles involve 5 finalist pipelines, 3 database slices, 3 cross-validated folds.

1.3 Robustness

The Venture Capital (VC) industry is concerned that predictive models trained on historical data will not predict future trends and activity. This has been identified as a key barrier to the adoption of automated systems by the VC industry [1]. Therefore, it is critical that our system is shown to be robust in its performance with respect to time so investors can rely on its predictions.

1.3.1 Training Set Date

We generated three models from datasets created from our training database from each year of 2012-2014 for forecast windows of 2 years (i.e. [2012, 2014], [2013,

2015], and [2014, 2016]) and evaluated each model against a dataset created from our test database (i.e. [2015, 2017]). We expected that if the factors that predict startup investment success through time are consistent, we would observe little difference between the performance and characteristics of these models.

Figure 1.4 shows the standard deviations of models trained on dataset slices from different years, against key evaluation metrics. We grouped by forecast windows as later dataset slices cannot be tested with long forecast windows which would skew results along this dimension. Variance across metrics is low, with more variance over shorter forecast windows.

We explored the feature weights for each model in Figure 1.5. While there are some slight differences, the general trend is similar across all models. We will discuss the distribution of these feature weights in more detail in a following section.

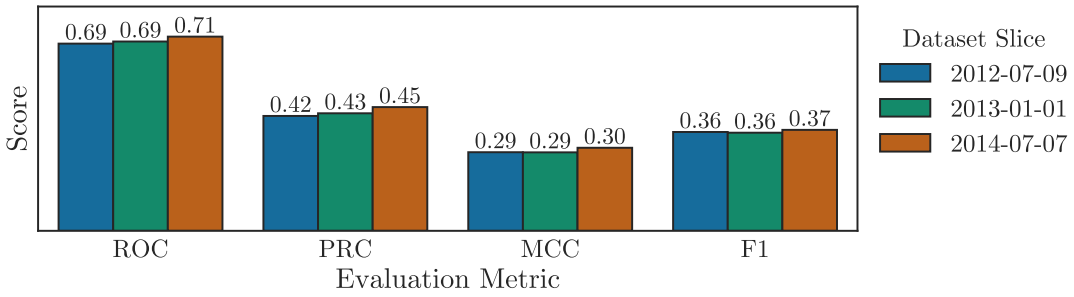

Figure 1.4: Performance variation by slice date.

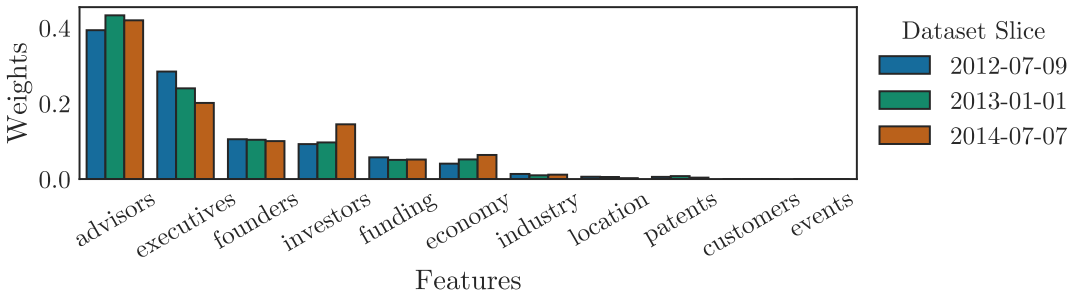

Figure 1.5: Feature weight variation by slice date.

1.3.2 Training Set Size

Learning curves allow us to evaluate how the bias and variance of a classification technique varies with respect to the amount of training data available. We

investigated learning curves for our system to determine whether our system's performance has potential to improve as its data sources grow. We sampled our training sets five times at different fractional rates. The rate of convergence (or divergence) of our training and cross-validation curves implies whether our classification pipeline is over- or under-fitting our data for various sizes.

Figure 1.6 shows the learning curves for forecast windows of 2-4 years against a combined target outcome and companies from all developmental stages. The maximum number of training examples is negatively related to the length of the forecast window because newer datasets have more examples. For a forecast window of four years the curves have converged, whereas for shorter forecast windows there still seems to be some benefit to additional training examples. Much of the testing score improvement comes in the first 20,000 training examples, which suggests this pipeline is approaching optimal performance.

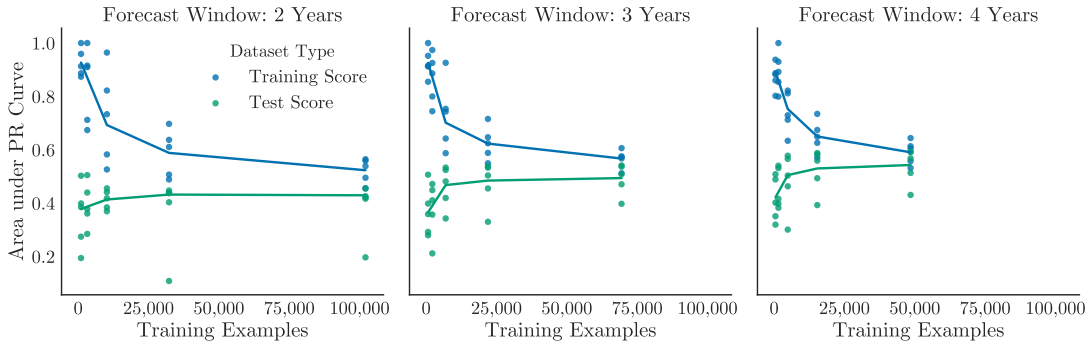

Figure 1.6: Learning curves by forecast window.

Figure 1.7 shows the learning curves as fitted independently to companies of different developmental stages, for a forecast window of four years and a combined target outcome. We observe significant variance in the learning curves for different developmental stages. The learning curves of Series B, C and D+ imply that more training examples will improve the performance of these models. Pre-Seed, Seed and Series A have converged or are at near-convergence at relatively low scores, and probably require more features or more complicated classification algorithms (e.g. Artificial Neural Networks) to improve their performance further. This is a reasonable observation given that companies in these earlier developmental stages have the most missing features in our dataset.

Observing When our learning curves are split by components of this target outcome, we see that the efficiency of our system varies, as shown in Figure 1.8. Predicting whether a company raises an extra round is the least data-intensive outcome, as it converges quickly. In comparison, predicting company exits does

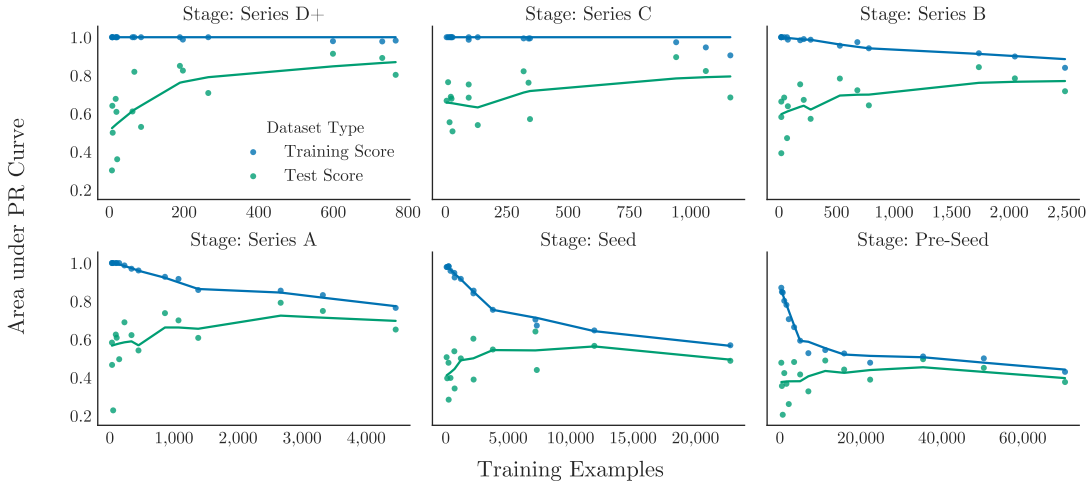

Figure 1.7: Learning curves by developmental stage.

not converge. Our model has most difficulty predicting Initial Public Offerings (IPO) which are rare events in our dataset.

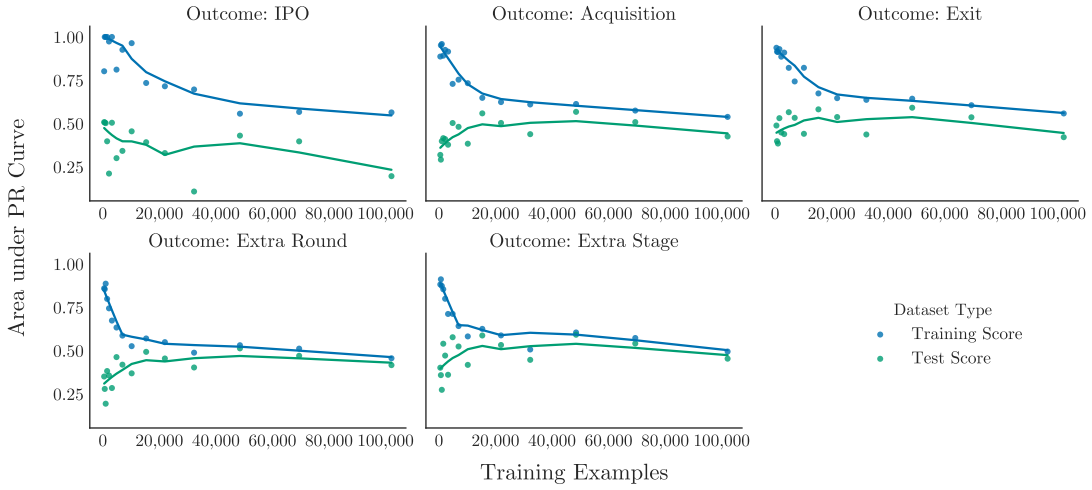

Figure 1.8: Learning curves by target outcome.

1.4 Versatility

Our system must be consistently accurate at identifying a variety of high-potential investment candidates. We evaluated the systems' versatility based on its ability to predict over different forecast windows (e.g. 2-4 years), for target companies at

different developmental stages (e.g. Seed, Series A etc.), and for different target outcomes (e.g. predicting additional funding rounds, being acquired, having an Initial Public Offering (IPO), or some combination thereof).

1.4.1 Forecast Windows

A forecast window is the period of time between when a prediction is made and when that prediction is evaluated (i.e. a prediction made in 2014 on whether a company would exit by 2017 is a forecast window of 3 years.) The Venture Capital (VC) industry raises funds with fixed investment horizons (3-6 years) [2], so time to payback is a key component of VC investment decision-making and portfolio management. It is important we understand how the models and predictions produced by a VC investment screening system varies with respect to the length of these forecast windows.

Figure 1.9 shows model performance across a range of metrics, grouped by forecast window. We observe little difference in Area under the Receiver Operating Characteristic (ROC) curve across the forecast windows. However, across all three other metrics, there is a positive relationship between length of forecast window and model performance. The F1 Score shows the greatest improvement in performance over time (52.7%), compared to Area under the Precision-Recall (PR) curve (34.1%) and Matthews Correlation Coefficient (MCC) (11.6%).

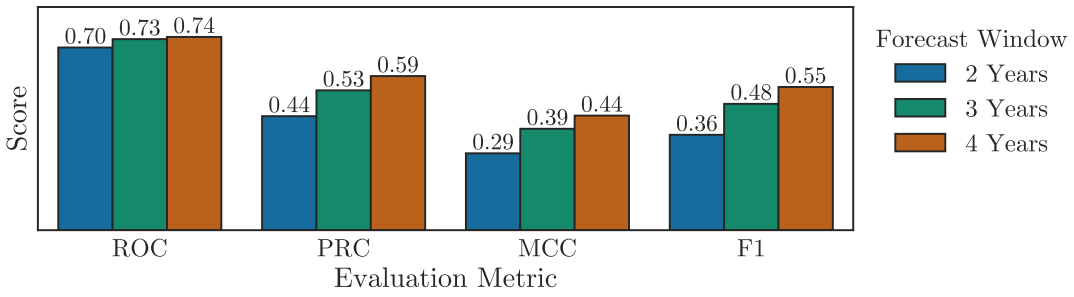

Figure 1.9: Performance by forecast window.

Figure 1.10 shows the standardised weights of features grouped using the conceptual framework proposed earlier in this paper, grouped by forecast window. First, we discuss the baseline distribution and then examine the variation in weightings with respect to forecast window. Advisors are the best predictor of startup investment success. Executives and founders are also important factors, and round out measures of human capital. The quality of investors that invest in a startup (assessed by their prior investments) is found to be more important

than the quantum of investment raised by a startup. Local economy and industry factors are weak predictors, as are customers and social influence (in this case measured through participation at events). There is little difference between the weightings of each feature group with respect to forecast window. However, there are a few trends to point out: the importance of advisors increases over time, and the importance of executives and the broader economy decreases over time.

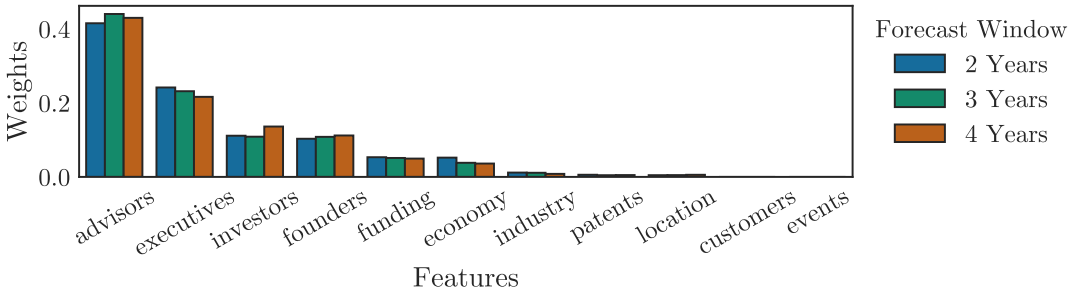

Figure 1.10: Feature weights by forecast window.

1.4.2 Development Stage

Startups can be classified into developmental stages based on their external funding milestones. These milestones not only signal a change in the resources available to a startup, but also their functions and objectives, and in turn the type of investors that are interested in them as investment opportunities. In Chapter **??** we mapped the companies in our dataset to their developmental stages. In the following section, we evaluated how the system models and predicts the outcomes of companies at different developmental stages.

Figure 1.11 shows F1 Scores grouped by developmental stage and fit method. First, we exmaine the baseline distribution and then the variation in performance by fit method. Model performance has a positive relationship with developmental stage. The only deviation from this relationship is for Series D+. To understand this discrepancy better, we split the datasets into their developmental stages and fit the model onto each of these sub-datasets individually. This results in a broad performance improvement. This method has the least impact on Pre-Seed and the greatest impact on Series D+ companies.

Figure 1.12 shows the standardised weights of features, grouped by developmental stage. While a similar trend to Figure 1.10 is observed, there is more varation in weights than was observed when grouped by forecast window. Advisors are more important to earlier stage companies than late stage companies,

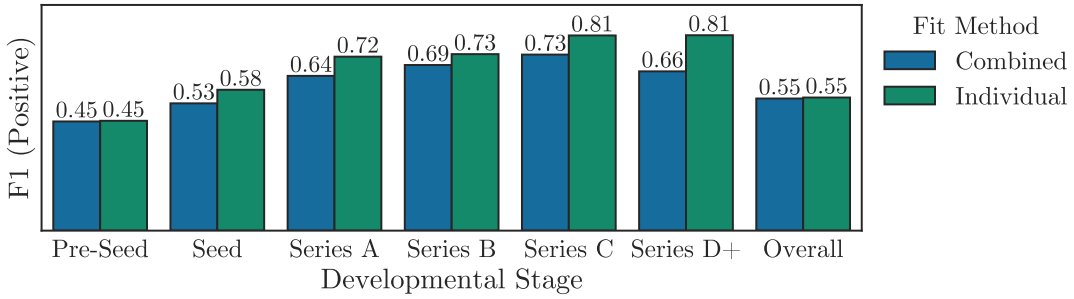

Figure 1.11: Performance by developmental stage.

investor track record and reputation becomes important as companies approach an exit (Series D+), executive and founder experience are important in pre-seed companies, as is broader economic outlook.

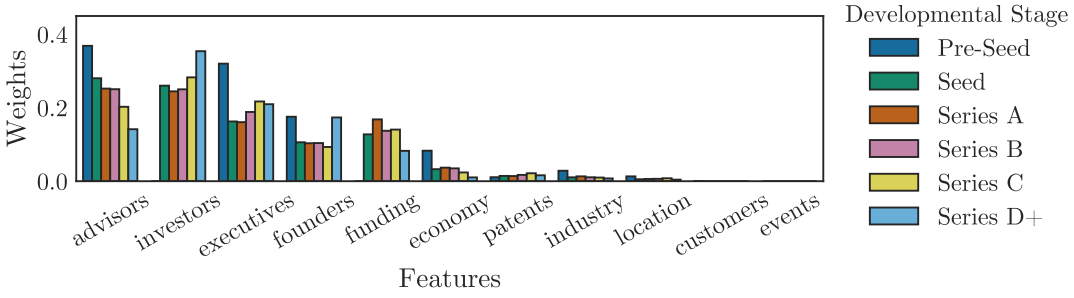

Figure 1.12: Feature weights by developmental stage.

1.4.3 Target Outcomes

Ultimately, VC firms seek rare investments that will return their invested funds many times over within an investment horizon of their fund (3-8 years). Funds are only returned to VC investors when startups have liquidity events (IPO, Acquisition). However, recently, many companies that are considered successful are delaying their liquidity events and seeking later-stage private funding (e.g. Uber). In this case, whether a company has raised additional funding rounds may be used as a proxy for investment success. Unless otherwise specified, we performed our previous analyses against our base target outcome, Extra Stage (i.e. whether a company raises an additional funding round, is aquired or has an IPO). In the following section, we explore whether the component outcomes (e.g. predicting IPOs) has an affect on our system's predictive power.

Figure 1.11 shows F1 Scores grouped by target outcome and forecast window. First, we examine the baseline distribution and then the variation in performance by forecast window. Our model is most accurate at predicting extra funding rounds and worst at predicting IPOs. As we observed in Figure 1.9, there is a positive relationship between length of forecast window and model performance. This relationship has a similar magnitude across all target outcomes except for IPOs which improve much more when the forecast window is increased from 2 to 3 years.

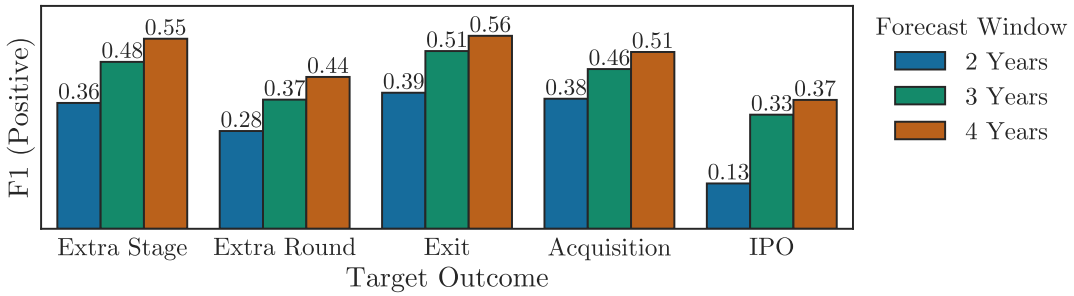

Figure 1.13: Performance by target outcome.

Figure 1.14 shows the standardised feature weight distribution, grouped by target outcome. Models of target outcomes produce considerable variance in feature weights. Exit and Acquisition have similar feature weights. Investors, Executives and Founders are key features for Exits and Acquisitions. In comparison, IPOs have more weighting towards Funding, Advisors and the Broader Economy. Extra Round is most strongly related to Investors and Funding and Extra Stage is most strongly related to Advisors.

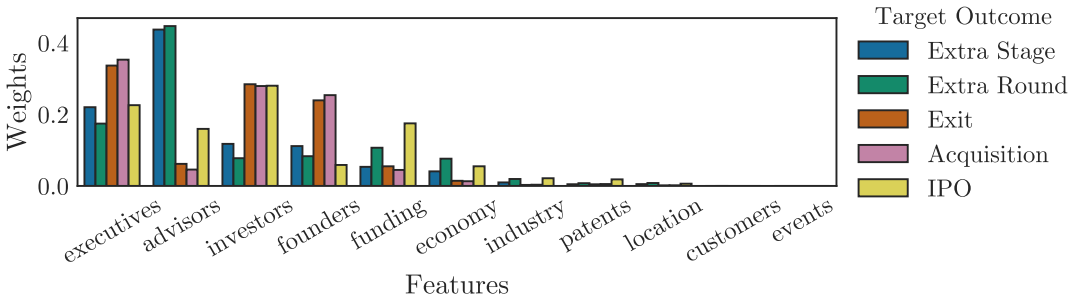

Figure 1.14: Feature weights by target outcome.

Bibliography

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

[2] Gompers, P. A. Optimal investment, monitoring, and the staging of venture capital. *The journal of finance* 50, 5 (1995), 1461–1489.