CHAPTER 1

Evaluation

We believe it is possible to produce a Venture Capital (VC) investment screening system that is efficient, robust and powerful. In Chapter ??, we described the development and structure of such a system. Our system identifies startup companies likely to receive additional funding or exit in a given forecast window. This system generates statistics and make recommendations that may assist VC firms to efficiently and effectively screen investment candidates. In this chapter, we evaluate models developed by our system against criteria of efficiency, robustness and predictive power.

We produced a classification pipeline optimised with respect to its robustness over time, and evaluated models produced by this pipeline against a held-out test dataset. This evaluation process is depicted in Figure 1.1. The pipeline is fit to a training dataset. The model is applied to a test feature vector to produce predictions. We score 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.

While Area under the Precision-Recall (PR) Curve were used to guide the development of our system during pipeline creation and selection, 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're 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. It should produce similar results to a macro-averaged F1 Score, incorporating the performance of both classes.

Firstly, we evaluated efficiency by exploring the learning curves of our classification techniques and whether there is sufficient data to produce reliable statistics. We also explored the time profile of our system and whether it is reasonable for use in industry, and would be likely to reduce the time currently taken to perform similar analyses. Secondly, we evaluated robustness by evaluating our models against multiple reverse-engineered historical datasets and measuring their variance. Thirdly, we evaluated the system's predictive power across different forecast windows, for startups at different stages of their development lifecycle, and for different potential target outcomes.

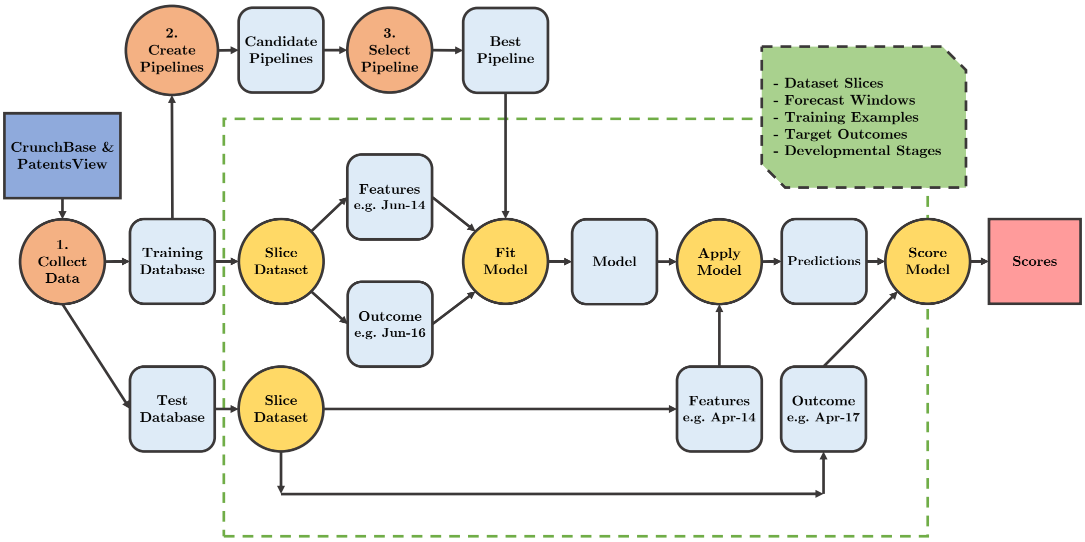

Figure 1.1: Pipeline evaluation overview.

1.1 Efficiency

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. By its nature, our automated system should be more efficient than these methods. In this section, we assess how efficient our system is – in terms of data consumed and time taken – and look at whether we can further improve its efficiency.

1.1.1 Dataset 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 classification pipeline to determine whether smaller samples could achieve similar predictive power and reduce the system's computational demand. We applied 10-fold stratified cross-validation to split our dataset into 10 subsets of different sizes which we used to train the estimator and produce training and test scores for each subset size. The rate of convergence of our training and cross-validation curves implies whether our classification pipeline is over- or under-fitting our data for various sizes allowing us to select an optimal sample size.

Figure 1.2 shows the learning curves for forecast windows of 2-4 years. 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 4 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 that this pipeline configuration is approaching peak performance. When the system is run in the future (with a larger dataset), the pipeline creation process may choose a classifier with less bias and more variance, like a Support Vector Machine (SVM) or Artificial Neural Network (ANN).

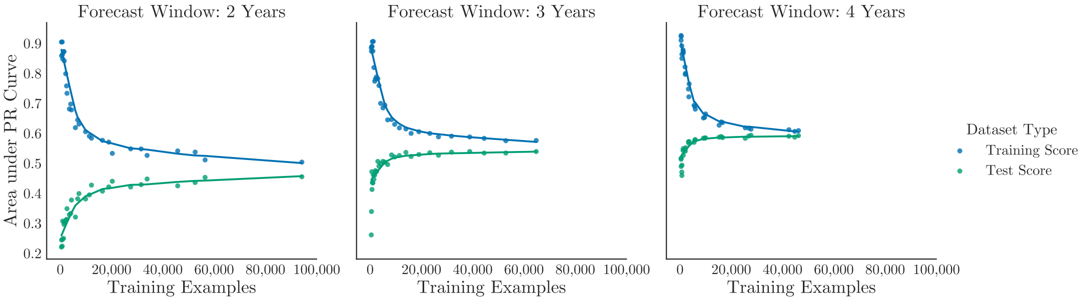

Figure 1.2: Learning curves by forecast window.

The plots in Figure 1.2 are evaluated against our base target outcome, which we term "Extra Stage" (i.e. whether a company raises an additional funding round, is aquired or has an IPO). 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 1.3. We observe that predicting whether a company raises an extra round is the least data-intensive outcome, as it converges rapidly even over a forecast window of 2 years. In comparison, predicting company exits does not converge, even over a forecast window of 4 years. Our model has most difficulty predicting IPO exits, which are rare events even in our large dataset. For these target outcomes, we would expect our system would benefit from a larger dataset. It should be noted, however, that 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 provide better performance.

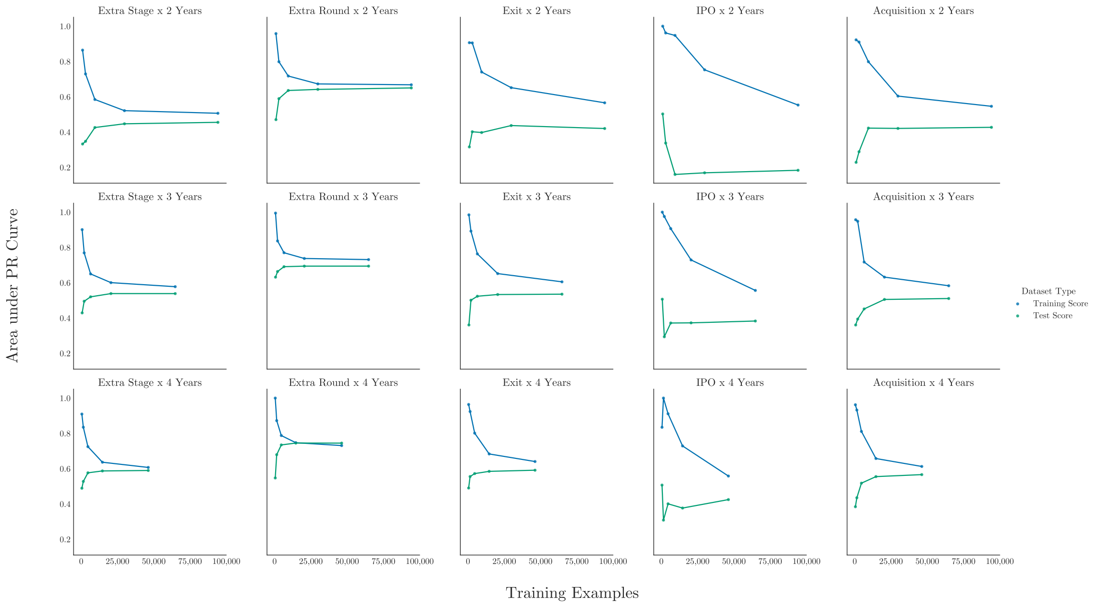

Figure 1.3: Learning curves by target outcome (column) and forecast window (row).

1.1.2 Time Profile

Unlike other forms of finance, like equity or derivatives trading, VC operates on a much longer timeframe – deals typically 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 significantly decrease the time taken to generate investment opportunities. We investigated the time profile of our system to determine whether it is practical for use in the VC industry.

An indicative time profile of the system is shown in Table 1.1. At the highest-level, this configuration of the program takes approximately 46 hours to complete on a modern desktop PC. When we further break this time down by system component, it's clear that 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 particularly long time because, in this case, it also involves generating learning curves for reporting, which is another cross-validated process. However, when placed into production, this component could be run infrequently - perhaps once per year - to ensure that 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.

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.

1.2 Robustness

The Venture Capital (VC) industry is concerned that predictive models trained on historical data will not accurately 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.

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 coefficient of variation of models trained on dataset slices from different years, against key evaluation metrics. The coefficient of variance is the ratio of the biased standard deviation to the mean. This produces a standard measure of variance, so different evaluation metrics are comparable. We have also grouped by forecast windows as later dataset slices cannot be tested with long forecast windows which skews results along this dimension. Variance across all metrics is very low, with slightly more variance over shorter forecast windows, as one would expect.

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

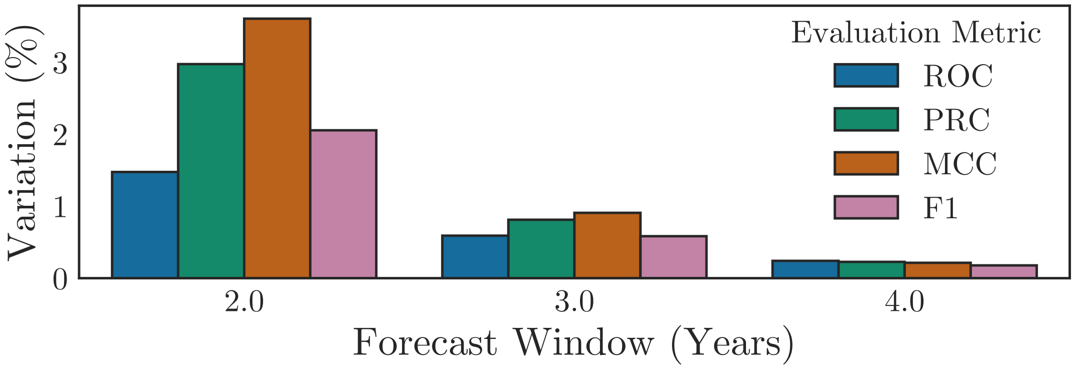

Figure 1.4: Performance variation by slice date.

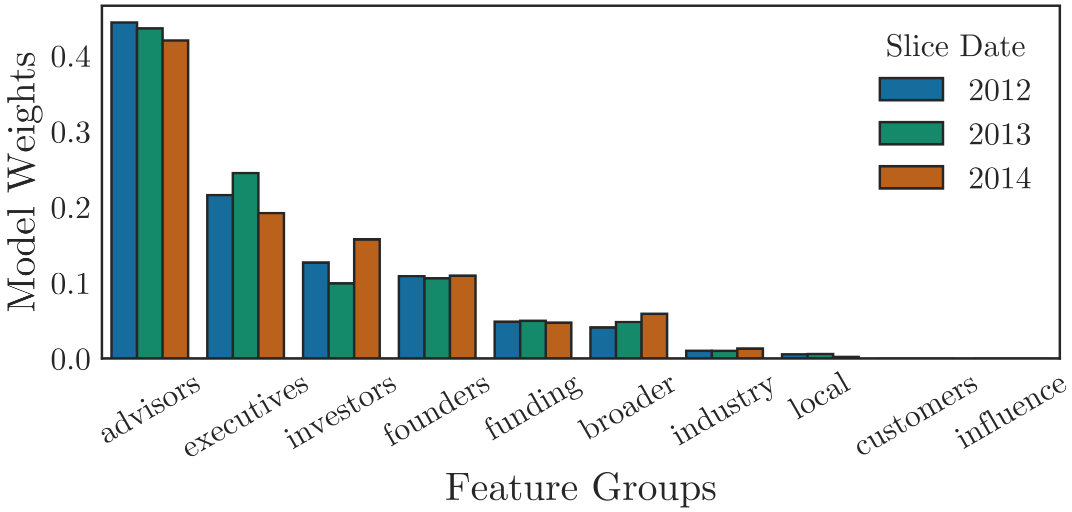

Figure 1.5: Feature weights by slice date.

1.3 Predictive Power

The system must be consistently accurate at identifying a variety of high-potential investment candidates. We evaluated the systems' predictive power 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 IPO, or some combination thereof).

1.3.1 Baseline Analysis

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

First, we looked at 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 difference between the slice that provides our features and the slice that provides our outcome. We combined pair-wise datasets of each year from 2012-2016 inclusive and explored the proportion of companies that raised additional funding or exited.

Figure 1.6 shows how company outcome varies with respect to the forecast window (time between the observed features and the measured outcome). Intuitively, we see a positive relationship between length of forecast window and company outcome. In particular, very few companies appear to have exited or raised funds over a period of less than 2 years so we will focus our experimentation on forecast windows of 2-4 years.

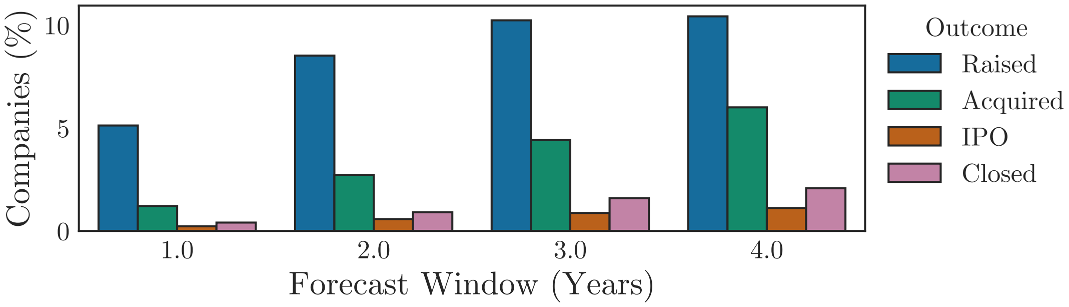

Figure 1.6: Outcomes by forecast window.

We also looked at how company outcomes vary with respect to development stage, shown in Figure 1.7. 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 suggested that in our experimentation we should investigate how our system predicts each stage independently, as well as in aggregate, as we do in a following section.

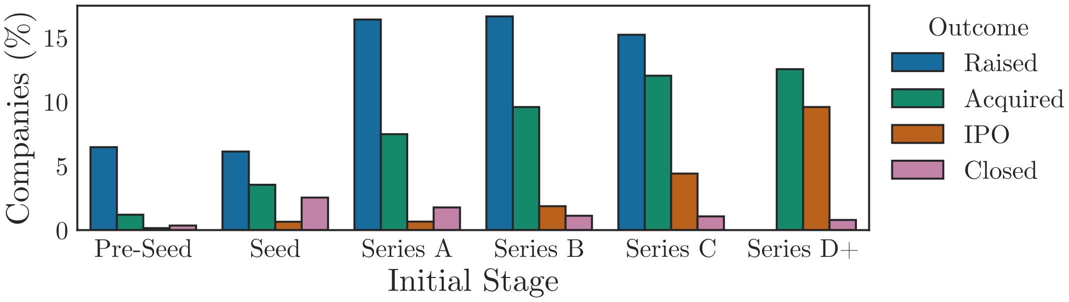

Figure 1.7: Outcomes by developmental stage.

1.3.2 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 (generally 3–8 years), 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 1.8 shows model performance across a range of metrics, grouped by forecast window. As discussed previously, we do not expect Area under the Receiver Operating Characteristic (ROC) curve to accurately reflect the performance of our model because it is not sensitive to our bias towards the positive class. We see that there is very little difference in Area under the ROC curve across the forecast windows. However, across all three other metrics, there is a clear positive relationship between length of forecast window and model performance. In particular, 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%), which probably reflects that our F1 Score here purely captures the positive class.

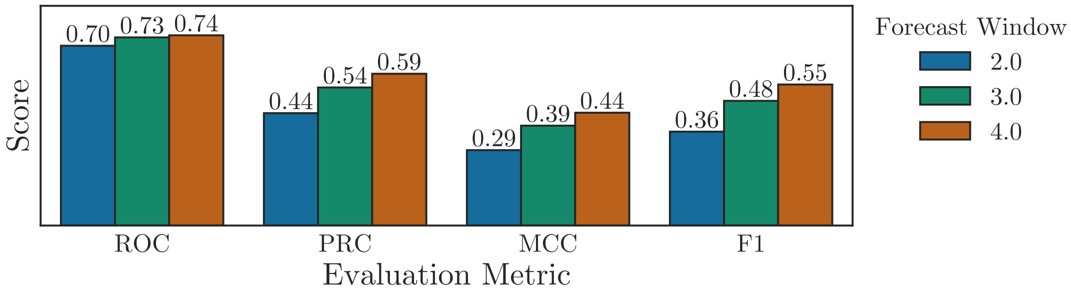

Figure 1.8: Performance by forecast window.

Figure 1.9 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, somewhat surprisingly, are the best predictor of startup investment success. This may reflect that exceptional startups are more effective at attracting influential advisors. 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). These factors are sparsely represented in the CrunchBase database.

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.

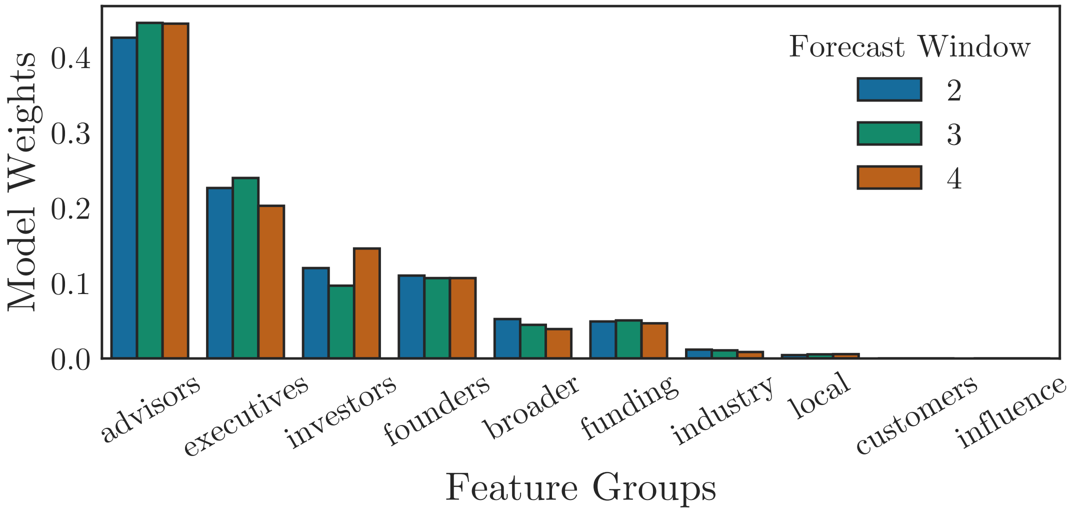

Figure 1.9: Feature weights by forecast window.

1.3.3 Development Stage

Startups can be broadly classified into developmental stages by virtue of 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 1.10 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. This may be a product of later stage companies having more complete feature vectors. The only deviation from this relationship is for Series D+. This may be because the model is only predicting exits at this stage. 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. Pre-Seed companies make up most of our original dataset and we see the smallest difference between methods for this stage. However, for Series D+ we see a significant difference in performance, which may suggest that the features that predict Series D+ performance are different to in earlier stages.

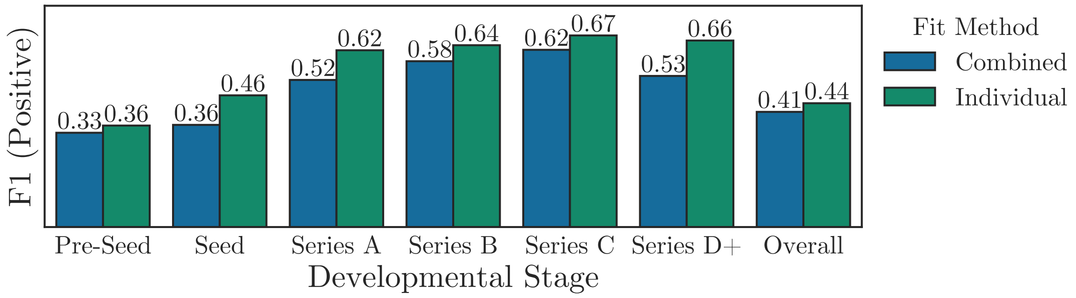

Figure 1.10: Performance by developmental stage.

Figure 1.11 shows the standardised weights of features, grouped by developmental stage. While a similar trend to Figure 1.9 is clear, 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, investor track record and reputation becomes important as companies approach an exit (Series D+), executive and founder experience are very important in pre-seed companies, as is the broader economic outlook.

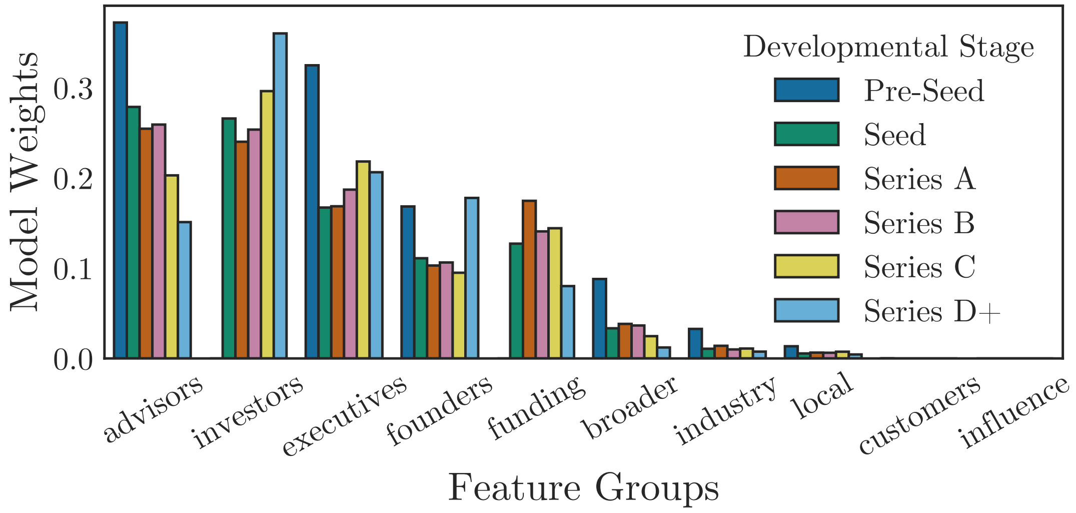

Figure 1.11: Feature weights by developmental stage.

1.3.4 Target Outcomes

Ultimately, VC firms seek rare investments that will return their invested funds many times over within an investment horizon of their fund (typically 3-8 years). Funds are generally only returned to VC investors when startups have liquidity events (IPO, Acquisition). However, particularly recently, many companies that are considered highly successful are delaying their liquidity events (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, which we term “Extra Stage” (i.e. whether a company raises an additional funding round, is aquired or has an IPO). In the following section, we will explore whether the component outcomes (e.g. predicting IPOs) will have an affect on our system’s predictive power.

Figure 1.10 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 performs badly at predicting IPOs, though these are rare events in our dataset. As we observed in Figure 1.8, there is a clear positive relationship between length of forecast window and model performance. We observe fairly similar performance improvements across the target outcomes except for in the case of IPOs, which improve dramatically from a forecast window of more than 2 years. This may be because there are non-performance related factors that affect IPO timing, so predicting the timing of an IPO is difficult for our model to achieve.

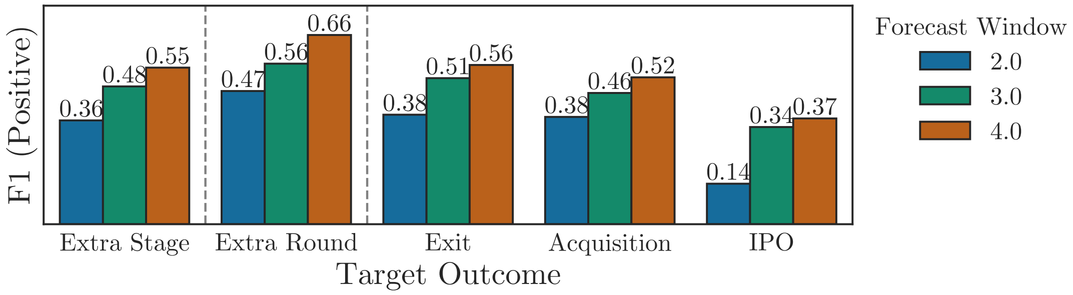

Figure 1.12: Performance by target outcome.

Figure 1.13 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, probably because Acquisitions make up a large proportion of Exits in our database. 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. Extra Stage is, perhaps surprisingly, most strongly related to Advisors.

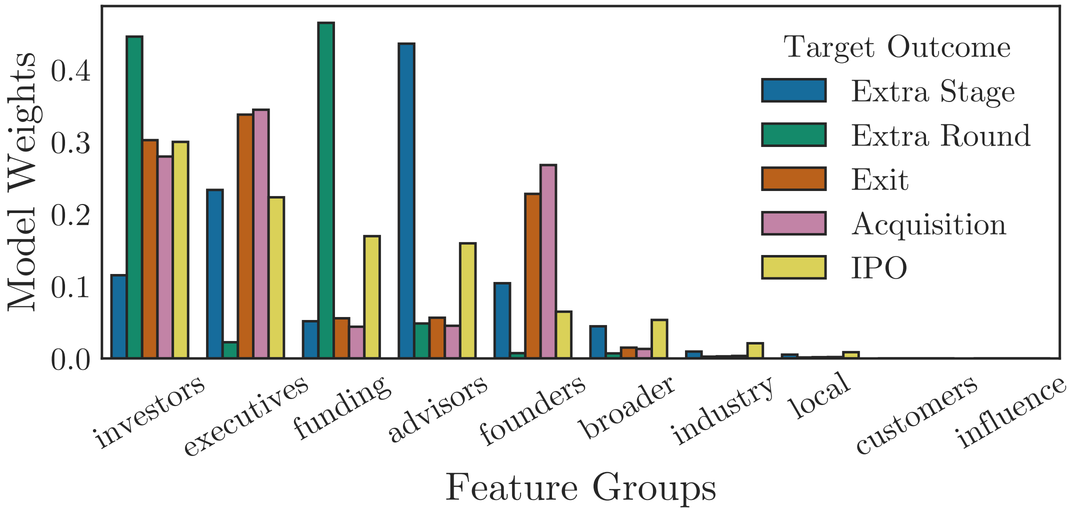

Figure 1.13: Feature weights by target outcome.

Bibliography

[1] Stone, T. R. "Computational analytics for venture finance". PhD thesis. UCL (University College London), 2014.