Baron Wilton
Kaavash Bahl
Meghana Harish
Shiyu Qian
Shiyue Ma

Section A
Team A-7

Startup Success and Evaluation

Industry: Entrepreneurship and Startups

Business Understanding

For the purpose of this assignment, we will be analyzing the startups dataset, with the success or failure of the start-up as a binary dependent variable, to determine which metrics most critically impact the likelihood of success for start-up companies across various industries. This project will generate meaningful insights for entrepreneurs looking to increase their chances of success and to create important key performance indicators (KPI's) when building their businesses.

In order to gain a better understanding of our dataset, we have summarized some causes for start-up failure:

Problems	*Solutions*
No market needs	Cannot solve with our current dataset
Fail to attract sustainable funds	Acquire more long-term investors
Get outcompeted	Benchmark on competitor's skills
Not the right team	Add advisers and executive team members with desired skillsets that relate to potentially higher success rates
Ignore customers	Focus on impact of subscription services, mobile apps, and the industries serviced

At the end of this project, our goal is to be able to locate the most crucial factors among various backgrounds of founders and structures of companies and to help future startups to decide what are the metrics that they can improve in order to.

Data Understanding

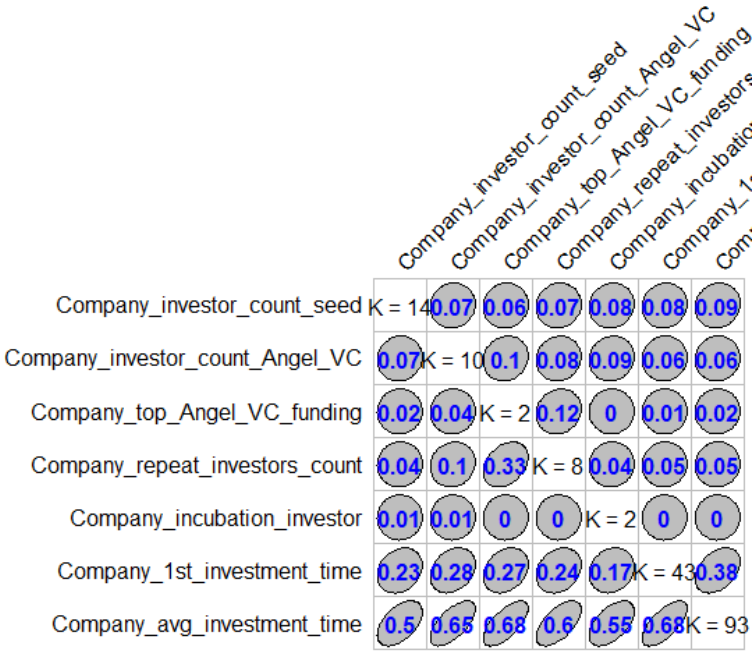
Our dataset contains 51 columns and 234 observations. The target variable is Success (1) or Fail (0) for the startups, and the independent variables cover information about various aspects of startups, which can be further divided into two major attributes: Company and Founder. Under the Company, we have variables providing general information about startups such as location, number of co-founders, business model, number of industries serviced, and financial information such as number of investors in different stages and investor level. For the Founder, variables include startup founders' background education, previous work experience, industry exposure, publications, and their skills for business, entrepreneurship, leadership and other perspectives.

Data Preparation and Cleaning

To perform better analysis, we first decided to drop some columns in our dataset since there are 51 variables. The followings are the columns that were dropped:

- CAX_ID: Unique ID number for each startup company
- Founder_experience: Average experience of founders. We decided to drop this one because it is an ordinal variable explaining the average experience of founders from low to high, and it is highly correlated with other variables

- Founders_skills_score: Overall skills for founders. We dropped this column because the information is too general and there are detailed founders' skills scores in different function given by other columns
- Company_avg_investment_time: Average time in months between multiple rounds of investment. We dropped this column due to multicollinearity shown in the graph below.

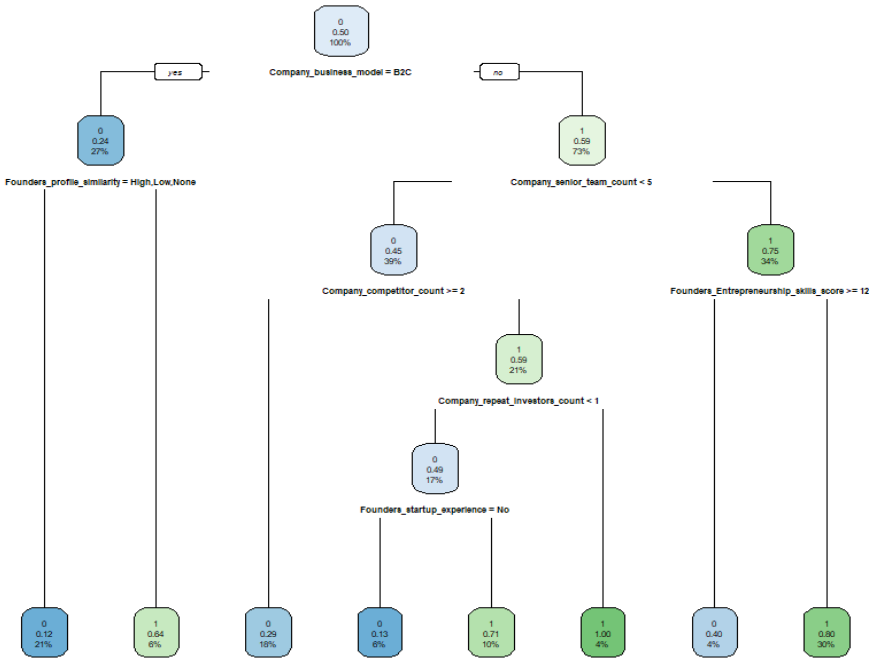


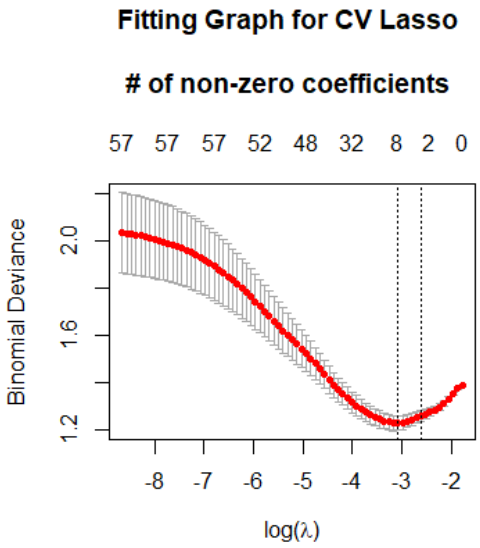
Before modeling, we tried to perform some unsupervised learning to help us understand the dataset and find segments. However, having too many binary and categorical variables in the dataset makes k-means not a good option, because calculating the mean is not meaningful to categorical variables. And with a total number of 45 explanatory variables also makes it really hard to interpret the results of 58 principal components. So we decided to rely on models to help us select the variables that have the most significant impact on the success of a startup first.

Modeling

We performed the following models to determine which variables in our dataset are most impactful in predicting successful startups:

- **Logistic regression:** using forward selection helped us selected 10 variables, and the AIC of the model is 272.
- **Logistic regression with interaction:** adding interaction into the logistic regression helped reduce the AIC to 122.
- **Classification Tree:** splits the data into 7 nodes resulting in success(1)/failure(0) .





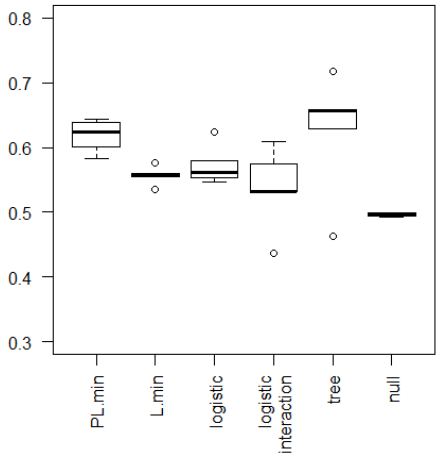
- **Lasso & Post lasso:** using the rule of λ=min selected 8 variables, λ=1se rule returned 4, and λ=theory rule only returned 2. So, we decided to choose λ=min rule among these three.

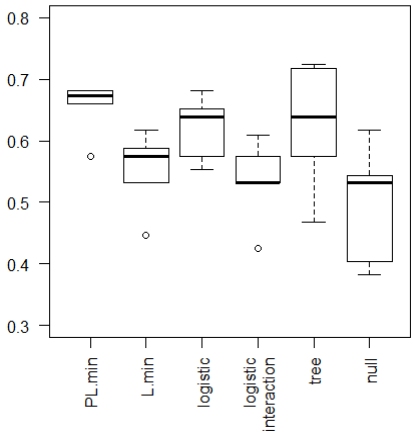
From all of our models, we can tell that Company_business_model and Company_senior_team count are the two variables that have the most significant impact on the success of a startup.

Evaluation

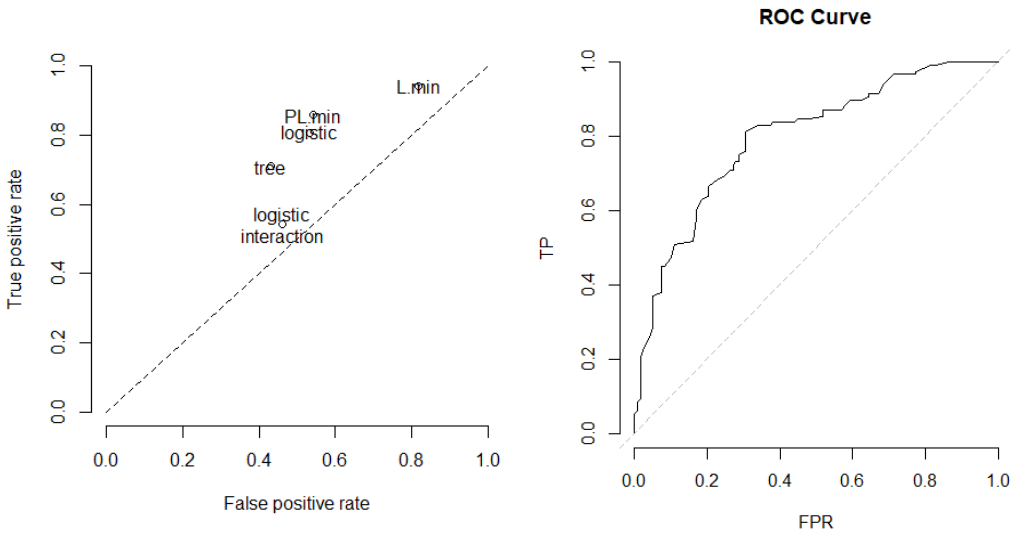
We used 5-fold cross validation to compare the out of sample performance of all our models. The baseline we used is the null model, which predicts that the startups have an average success rate of 49.5%. The 2 performance measure matrices we used are:

1. 1-mean(|Y-P(Success|X)|)
2. Accuracy

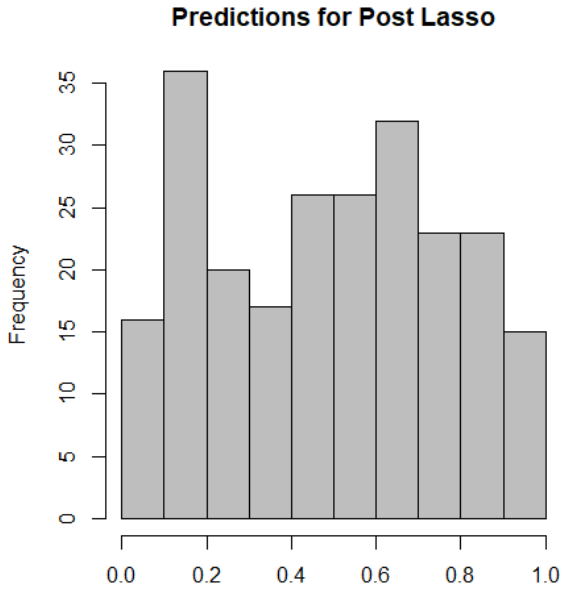




From the boxplots above, we can observe that among all of our models, post lasso using rule λ=min has the best OOS performance based on both of our performance measure matrices. According to the second graph which gives us OOS accuracy metrics, post lasso yields the highest accuracy with the lowest variance based on the range. And logistic regression with interaction has the worst OOS performance from both graphs, which can be explained by model overfitting. What's more, although classification has the highest average performance in the first graph, its prediction has the largest variance. Therefore, we used post lasso for our further computations.



Using a threshold value of 0.3, we calculated the TPR and FPR of the models to further confirm that post lasso was the way to go. Using the predictions from post lasso, we computed the confusion matrices for various threshold values to plot the ROC curve. The ROC curve gave us a large AUC which means that there is a higher chance for our post lasso model to be able to distinguish between positive and negative classes, hence providing us with more accurate predictions.



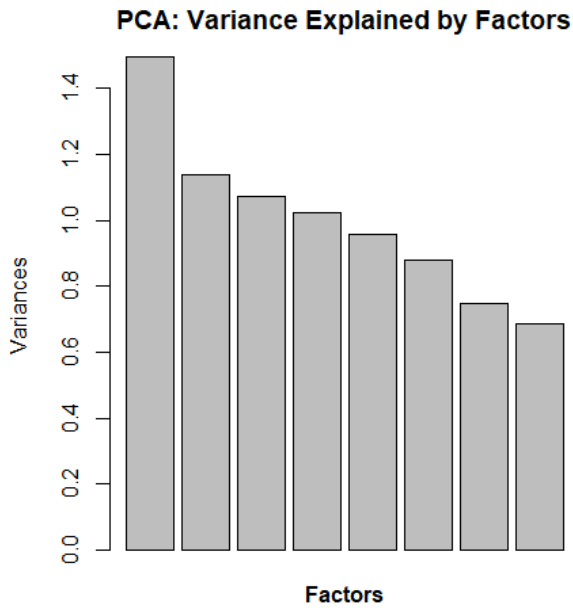
We also plotted the frequency of the predictions obtained from running post lasso to check for any abnormalities or outliers. We did not come across any such values and verified that post lasso would result in reliable predictions.

Deployment

After comparing each model, we ran a principal component analysis (PCA) using the 8 variables provided by the post lasso model. These variables include:

- Company_senior_team_count
- Company_business_model
- Founders_industry_exposure
- Founders_education
- Founders_publications
- Company_competitor_count
- Company_crowdfunding
- Company_analytics_score

Our PCA returned 8 principal components, and each principal component's characteristics are shown in the table below. To simplify the clustering process, we used the weighted score from prediction to determine the principal component that each company was most similar to. By default, each company is comprised of a combination of each principal component, but these weighted scores highlight which component is most salient for each company. Then, we calculated the mean prediction of our dependent variable for each principal component - giving us our "success rate" that can be found in the table below. Hence, back to our original business purpose, entrepreneurs can align themselves closer to the characteristics found in the more frequently successful principal components.



From the PCA histogram above, we generated eight PCs. PC1 through PC6 explain the variance of up to 82.1% of our data hence we dropped PC7 and PC8 from consideration. Hence, we have the following matrix:

PC	Average success Rate	Size	Characteristics	Potential Company Type
1	78.4%	51	• Majority of B2B business model • More senior team members • High analytical skills • Famous founders with many publications	Companies with high research output with experienced industry professionals- Healthcare Startups
2	55.6%	27	• Small companies in a highly competing industry • Low industry exposure of funders • High analytical skills • Supplied by crowdfunds	Wearable Tech Startups
3	24.0%	25	• Small B2B companies without	Online wholesaler

			crowdfunding ● Average performance on everything else	
4	41.2%	68	● Small B2C companies founded mostly by bachelors ● Low industry exposure of founders	E-commerce store
5	45.2%	31	● Middle size B2B companies in a niche market ● Low industry exposure of founders	Boutique consulting firms
6	40.6%	32	● Middle size B2C companies ● Medium industry exposure of funders ● High education level of founders	Tech startups

Recommendations

From this PCA, we recommend the following:

- Regardless of business model, crowdfunding is one of the most important activities entrepreneurs should strive for at early stages of business development.
- For B2B startups, a large senior team with industry background and skills rooted in analytics will help the firm succeed. This is important to build credibility amongst business clients who would value these attributes in the companies they are working with, especially if the firm is a startup in the space.
- B2C startups are much harder to successfully operate, and thus do not frequently succeed. However, entrepreneurs in the space will want to surround themselves with a large senior team, full of domain-specific experience and skill sets.
- Firms more frequently fail when they merely focus on higher academics and research but lack the requisite skills to run the business according to what the market demands.

Ironically, they also can fail if they have low numbers of competitors, meaning that they may be creating revolutionary new products or services that do not have any current companies to benchmark practices off of.

These models and principal components can be used as a benchmark to help entrepreneurs in the future to evaluate their startup performance and decide what aspects of the company that they should mostly invest in to increase their success rates. Entrepreneurship is full of risks. Our project aims to provide a bit of clarity in which aspects of new businesses entrepreneurs should focus on to reduce their risk of failure. Future research can be performed on datasets regarding long-term information on startups to determine the duration and magnitude of these firms' successes so entrepreneurs can not just establish a solid business, but also keep it profitable and running for years to come.