Vela Partners Micro-internship

Report

Start-up Success Prediction Tool

Aldaïr A. F. M. Petronilia

December 17, 2021

Abstract

In this short note, we explore a data set of success and unsuccessful companies. We attempt to construct features which produce a signal for start-up success. Lastly we train a neural network using these constructed features.

1 Introduction

The goal of this project is to capture features in a data-set such that a model may be employed to determine the success of a start-up. The difficulty in this project lies in the fact that the data provided is highly categorical. Simply vectorising the data and feeding it into a neural network model seemed infeasible due to only having approximately 50,000 data points to work with. The main strategy employed was to extract features from the data to reduce its dimensionality to train a model for prediction.

This document is organised as follows. In Section 2, we describe the stucture of the data that we have been given and what is meant by the term success. In Section 3, we perform various transformations on the datas-et to look for features that produce signals that a company may be successful. We employ statistical techniques in this section. Using the signals, we group the data to train a model. Lastly, in Section 4, we talk about the model we trained and the results obtained.

2 The Data Structure

The first step taken in trying to produce a predictive agent is to understand the structure of the data given. The main data set consists of two spreadsheets, one composed of *successful* companies and one composed of *unsuccessful* companies. The term *success* here is meant to imply that the company is a publicly traded company. The structure of each sheet is identical and presented in the table below.

Name	Founded Year	Country Code	City	...	Description
Name 1	Year 1	Code 1	City 1	...	Description 1
Name 2	Year 2	Code 2	City 2	...	Description 2
Name 3	Year 3	Code 3	City 3	...	Description 3
Name 4	Year 4	Code 4	City 4	...	Description 4
Name 5	Year 5	Code 5	City 5	...	Description 5
⋮	⋮	⋮	⋮	⋱	⋮
Name N	Year N	Code N	City N	...	Description N

Table 1: *Table showing the structure of the data-set*

The total list of columns provided and their values:

- **Name** - The name of the company.
- **Founded Year** - The year the company was founded.
- **Country Code** - The code of the country the company is registered in.
- **City** - The city where the company operates out of.
- **Category Group** - A broad category or domain that the company product is in.
- **Category** - A more specific category or domain the company's product is in.

- **University of Founders** - A list of universities that the founders have attended.
- **Degrees of founders** - The degrees held by the founders of the company.
- **Subject Degrees of Founders** - A study the founders did in university and obtained a degree in.
- **Gender of Founders** - The gender of the founders
- **City of Founders** - The city where the founder lives
- **Previous Companies of Founders** - List of the previous companies that the founders worked at.
- **Previouss Title of Founders** - Previous job titles of the founding team.
- **Investor Name** - A list of investors that have invested into the company .
- **Short Description** - A short description of the company.
- **Description** - A long description of the company.

3 Feature Generation

The primary challenge with the data provided is that is categorical. The question is how to convert this categorical data into actionable information.

3.1 Collation of the Data

The first step was to collate the data based on the number of founders. We counted the number of founders, gender of the founders and gender ratio of the founders. See Appendix A.1 for box plots that show the distribution of the data. Figure 6a shows that by having a larger number of founders, then you are more likely to have a successful start-up. This trend is more pronounced in Figure 6b. Figure Sub-Figure 6c and Figure 6 Sub-Figure 6d fail to display any actionable information mainly due that the founders of companies tend to be male dominated. Table 4 actually show that Pearson correlation between number of founders and success of a company is 0.1567. When looking at only at the number founds that are male, this correlation is slightly higher at 0.1699. This tells us that number of founders may produce a signal for the likely-hood of a company to be successful.

Next we studied the effect of investors on the success of the company. Two metrics were produce. First we computed the number of investors that each company had. Secondly, we computed the average investor success. Given an investor in the data-set, we computed the number of successful and unsuccessful companies that investor invested in. This allows you to compute an empirical probability of success. For start-up, we compute the average of the success of the investors that have invested in them. See Appendix A.2 for box plots that show the distribution of the data. Figure 7a shows that number of investors may produce a signal for the success of a company. This is intuitive as the more successful a company becomes, it will attract more investors. Figure 7b shows that there is a clear signal of a company being success depending on the success profile of the investors that have invested into the company. Table 5 shows the correlations between these two metrics and success of the start-up. There is a clear

strong correlation between the average success rate of the investors in the start-up and the startup becoming successful.

Next we studied how well educated the founding team is affects the success of a start-up. The rankings of the universities was computed using QS world university rankings 2022 data-set. We computed the best rank and score out of the universities attended by the founding team. We computed the worse rank and score out of the universities attended by the founding team. Lastly, we computed the average rank and score out of the universities attended by the founding team. See Appendix A.3 for box plots that show the distribution of the data. Figure 8a and 8d show a clear signal that if one of the founders went to a top ranked university, then there is a higher chance of the start-up being a success. Table 6 show that although the box plots may produce a signal there does not seem to be a clear correlation between the metrics produced and the success of a start-up. The figures in Appendix A.7 further solidify our analysis that the best rank of the university of the founders may be a signal for startup success as shown in Figure 12b. As the best rank increases, so does the probability of success.

Next, for lack of a better word, we computed some intelligence metrics. The data set included the degree obtained by the founders; this may have been a Doctorate, Masters, Bachelors, Associate, etc. The degree was mapped to a numerical value. Doctorate was mapped to 4, master's was mapped to 2, bachelors' was mapped to 1, associate's was mapped to 0.5 and all other cases was mapped to 0. See Appendix A.4 for box plots that show the distribution of the data. Figure 9a does not show much information due to the data still being categorical with the numerical mapping. Figure 9c shows that we may have a signal of success based on the average degree of the founders. Regardless of this signal in the distribution, Table 7 fails correlation between these produced metrics and success of the start-up. The figures in Appendix A.8 further solidify our analysis that the average degree level of the founders may be a signal for startup success as shown in Figure 12b. Average education level of the team increases, so does the probability of success.

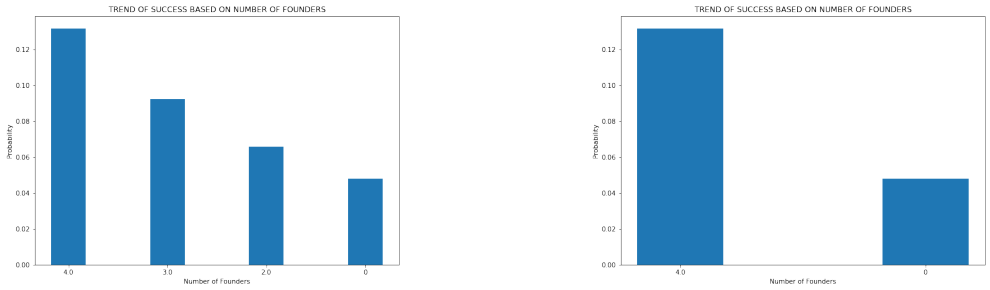
The data-set also included the previous job titles of the founding team. We counted the number of previous job titles of the founders. Furthermore we counted the number of jobs that was of a managerial level or leadership position within the company such as CEO, CFO, Chairs, etc. We also averaged these quantities against the number of members in the founding team as it is clear the larger the founding team, the more experience they would have. See Appendix A.5 for box plots that show the distribution of the data. Figure 10a and Figure 10c show a clear signal that as the experience of team increases whether that be managerial/ leadership experience or not, there is increase in the odds of being successful. Figure 10d and Figure 10d do not show any clear signals. Table 8 does not show that there is an apparent correlation between team experience and start-up success.

Lastly, we computed a connectedness factor. Looking at univeristies where the founders when to, counted how many went to the same university. We counted the largest size of the founders that went to the same university and divided this by the number of founders. This is what we denote as the connectedness factor of the founding team. See Appendix A.6 for box plots that show the distribution of the data. Figure **??** shows a clear signal that successful start-up have a founding team that is well connected. Although the box plot distribution shows a signal, there does not seem to be a correlation between connectedness and success of the start-up, Table 9

3.2 Data grouping

Using these metrics we have been able to reduce the sparsity of the data and see what may be a clear signals that lead to a successful start-up. To further reduce the sparsity of the data, it may be instructive to group the data together.

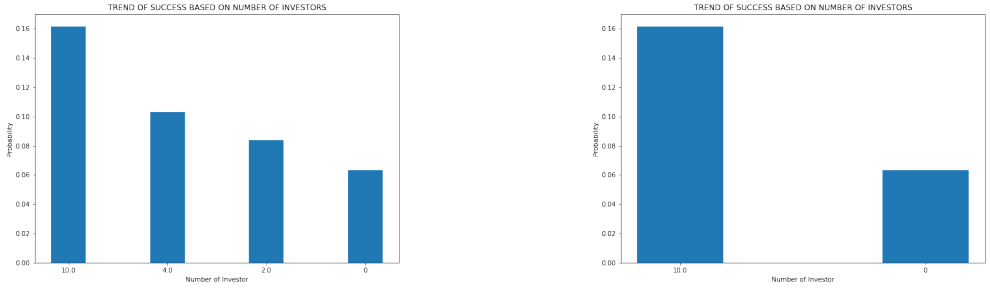
Proceeding in the same order as the previous section, we shall group the data of founders into 4 or 2 categories based on the empirical quantiles of the successful and unsuccessful data sets. Probabilities are calculated by looking at the companies that have at least as much founders for the value on the x-axis and counting how many of them are successful and unsuccessful to compute the probabilities.



(a) **(b)**

Figure 1: *Figure showing the probability of success based on having at least x amount of founders. The values on the $x-$axis is chosen by the quantiles of the empirical data-set*

Figure 2 show a clear trend that as the number of investors decrease, so does the probability of success of the start-up. This makes the signal for success. Next we shall group the number of investors of the startup into four/two groups and compute the probability of success.

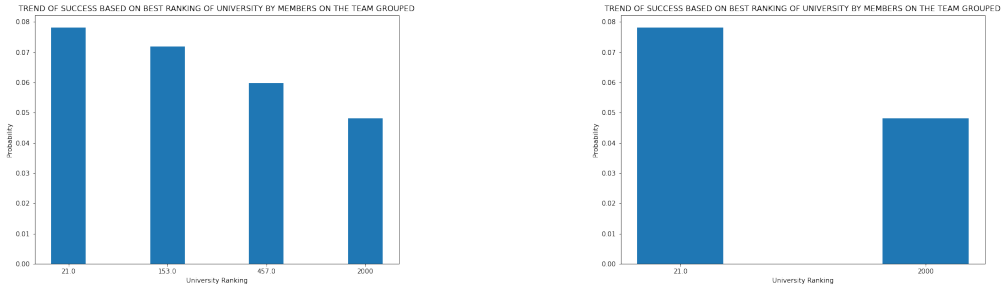


(a) **(b)**

Figure 2: *Figure showing the probability of success based on having at least x amount of investors. The values on the $x-$axis is chosen by the quantiles of the empirical data-set*

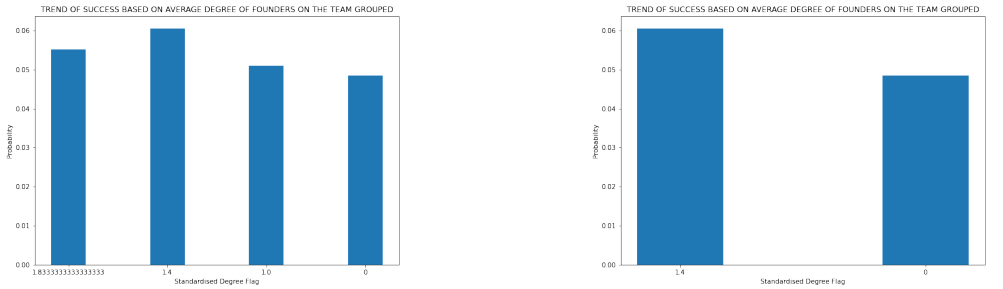
Similarly as before there is a clear signal that as the number of investors increase then so does the probability of success. Next we group the start-ups based on the best university of one of the founders on the team. We saw this was a signal of success in the previous section.

Figure 3a shows a clear trend that as the ranking of the university decreases, so does the probability of success. Next we group the start-ups based on the average education acquired by the founding team.



(a) **(b)**

Figure 3: *Figure showing the probability of success based on having a university ranking of at most x. The values on the $x-$axis is chosen by the quantiles of the empirical data-set*

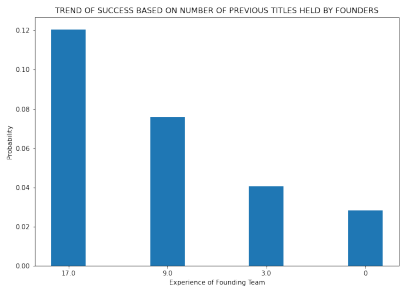


(a) **(b)**

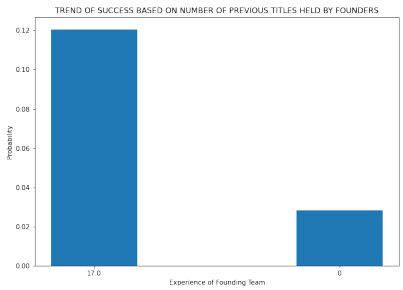
Figure 4: *Figure showing the probability of success based on having an average degree level of at least x. The values on the $x-$axis is chosen by the quantiles of the empirical data-set*

Figure 4a shows a clear phase transition where the probability of success drops as the average education level of the founders change. Lastly we look at the overall experience and managerial or higher experience of the founding team. The same trends as before appear in Figure 5.

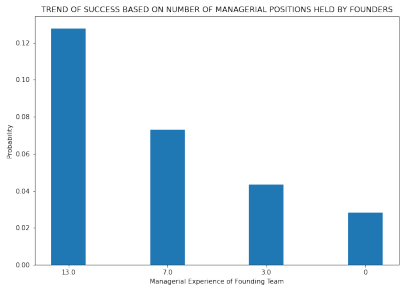
Lastly, we have not used the information relation to the category that the start-up operates in or the previous companies that the founders have worked at. To group this data we counted the popular category and category group that the start-ups operate in. We all counted the popular companies that founders have previously worked at. See Appendix A.9 and Appendix A.10 for the probabilities of success when operating in these domains. Start-ups were also groups with whether they were in this category or not.



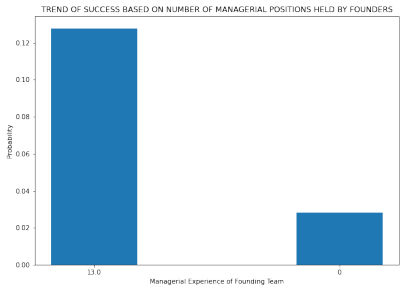
(a) *1a*



(b) *1b*



(c) *1a*



(d) *1b*

Figure 5: *Figure showing the probability of success based on having an experience level or managerial experience level of at least x. The values on the $x-$axis is chosen by the quantiles of the empirical data-set*

4 The Model

Using these features, we trained a deep neural network using the groupings generated. We observed that generally neural networks that used the groupings into four categories generally performed better when compared with neural networks that used the binary groupings.

We trained a neural network a deep neural network with 4 hidden layers with 128 neurons in each hidden layer and relu activation functions. For the output neuron, we used a sigmodial activation function. We trained the network based on the following grouped signals

- Number of founders
- Best Ranking of University of Founders
- Best Score of University of Founders
- Average Degree of the Founders
- Previous Experience of Founders
- Previous Managerial Experience of Founders
- Flag if Operates in Top 10 Category
- Flag if Operates in Top 10 Category Group
- Flag is A Founder Worked in a Top 10 Popular Company

- Number of investors
- Average Investors Success (Not Grouped)

We split the data 70/15/15 in training set, test set, validation set. The accuracy metric we used was Binary accuracy as we are interested in binary classification. We also tested the accuracy of predicting a successful company when given only successful companies.

Network	Number of Data Points (Total, successful)	Test Set Binary Accuracy	Accuracy of Successful Companies	Accuracy of Unsuccessful Companies
Network without Number of investors	$10305, 482$	95.6%	6.02%	99.8%
Network without Average Investors Success	$4687, 437$	94.5%	34.6%	98.7%
All features	$4687, 437$	99.9%	99.2%	99.2%

Table 2: *Table showing the accuracy of the trained neural network*

Table 2 shows that it is clear the the success rate of previous investors and number of investors are able to increase the predictive power of the model to predict the success of a start-up.

Lastly, we trained a model with the same parameters as above but with using the connectedness factor

Network	Number of Data Points (Total, successful)	Test Set Binary Accuracy	Accuracy of Successful Companies	Accuracy of Unsuccessful Companies
Network without Number of investors	$5685, 365$	95.0%	17.5%	99.2%
Network without Average Investors Success	$2940, 336$	95.6%	52.1%	98.2%
All features	$2940, 336$	100%	99.5%	99.5%

Table 3: *Table showing the accuracy of the trained neural network and employing the connectedness factor*

We see by employing a connectedness factor we are able to get higher accuracy of success prediction with networks that have less information. The high accuracy of the

last network is dubious and more data should be used to check how well such a network generalises.

Appendices

A Feature Generation Figures

A.1 Number of Founders

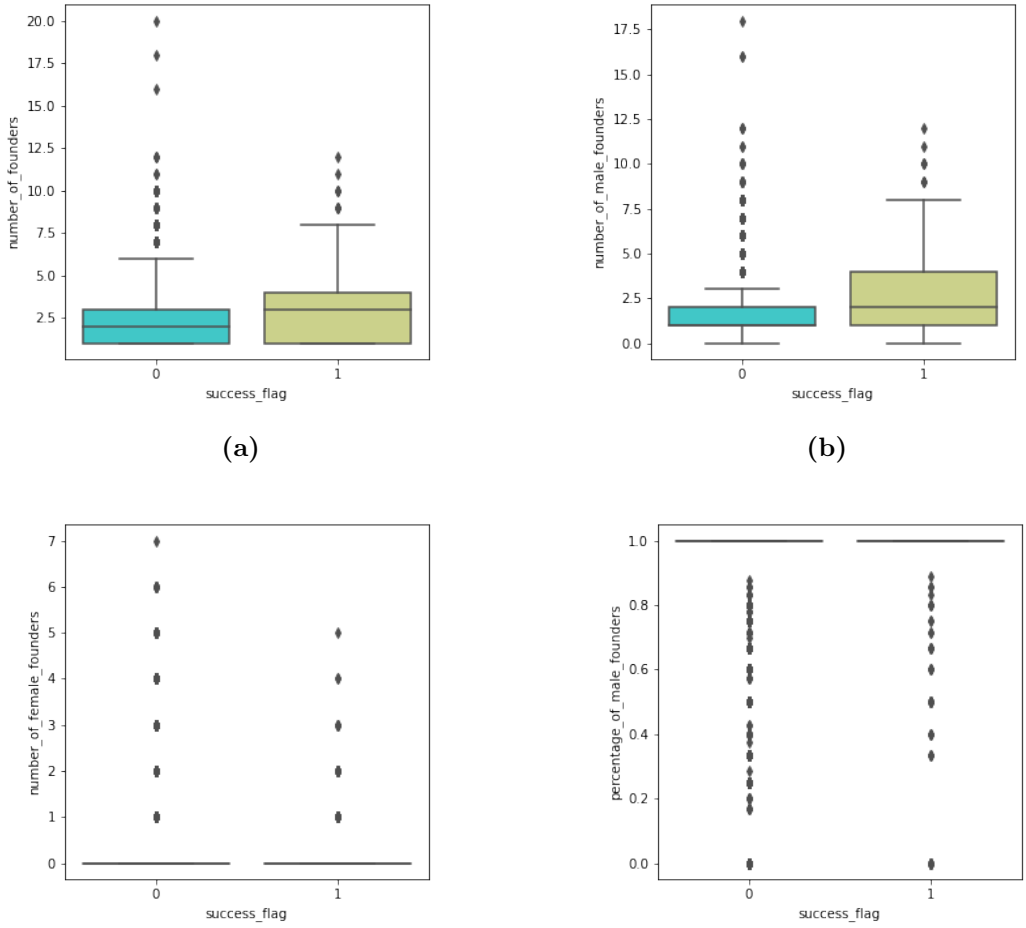
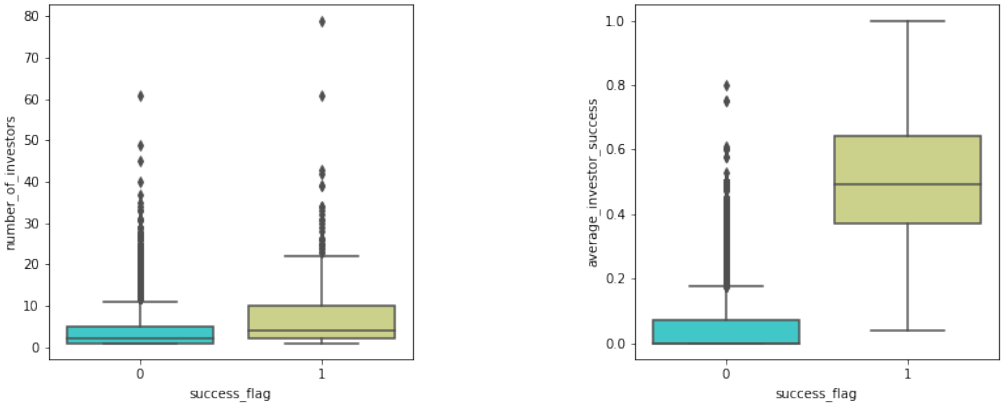


Figure 6: ***(a)*** *shows box plot distribution of success or failure of a company depending on the number of founders.* ***(b)*** *shows box plot distribution of success or failure of a company depending on the number of male founders.* ***(c)*** *shows box plot distribution of success or failure of a company depending on the number of female founders.* ***(d)*** *shows box plot distribution of success or failure of a company depending on gender ratio of founders*

Feature	Number of Founders	Number of Male Founders	Number of Female Founders	Percentage of Male Founders
Correlation	0.1567	0.1699	−0.0256	0.0502

Table 4: *Table showing the correlation between the features and the success of the start-up*

A.2 Number of Investors



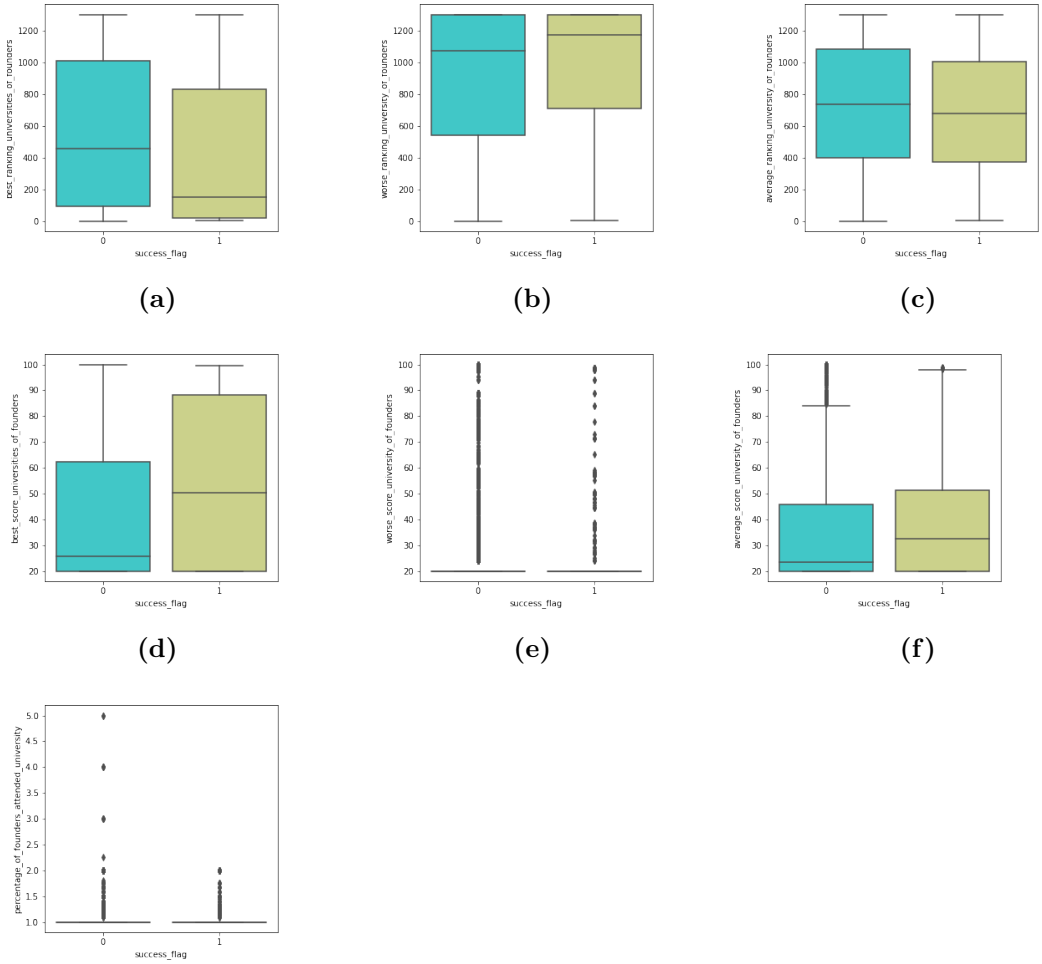
(a) **(b)**

Figure 7: ***(a)*** *shows box plot distribution of success or failure of a company depending on the number of investors.* ***(b)*** *shows box plot distribution of success or failure of a company depending on average success probability of the investors of the start-up*

Feature	Number of Investors	Average Success Rate of Investors
Correlation	0.1807	0.7589

Table 5: *Table showing the correlation between the features and the success of the start-up*

A.3 University of The Founders Ranking



(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g)

Figure 8: *(**a**) shows box plot distribution of success or failure of a company depending on the best ranking of the university of the founders.(**b**) shows box plot distribution of success or failure of a company depending on the worse ranking of the university of the founders.(**c**) shows box plot distribution of success or failure of a company depending on the average ranking of the university of the founders.(**d**) shows box plot distribution of success or failure of a company depending on the best score of the university of the founders.(**e**) shows box plot distribution of success or failure of a company depending on the worse score of the university of the founders.(**f**) shows box plot distribution of success or failure of a company depending on the average score of the university of the founders.(**g**) shows box plot distribution of success or failure of a company depending on the percentage of founders that have attended university.*

Feature	Best University Ranking	Worse University Ranking	Average University Ranking	Percentage Attended University
Correlation	−0.05858	0.02938	−0.02197	0.01697
Feature	Best University Score	Worse University Score	Average University Score	
Correlation	0.07831	−0.01708	0.03489	

Table 6: *Table showing the correlation between the features and the success of the start-up*

A.4 Intelligence Metrics

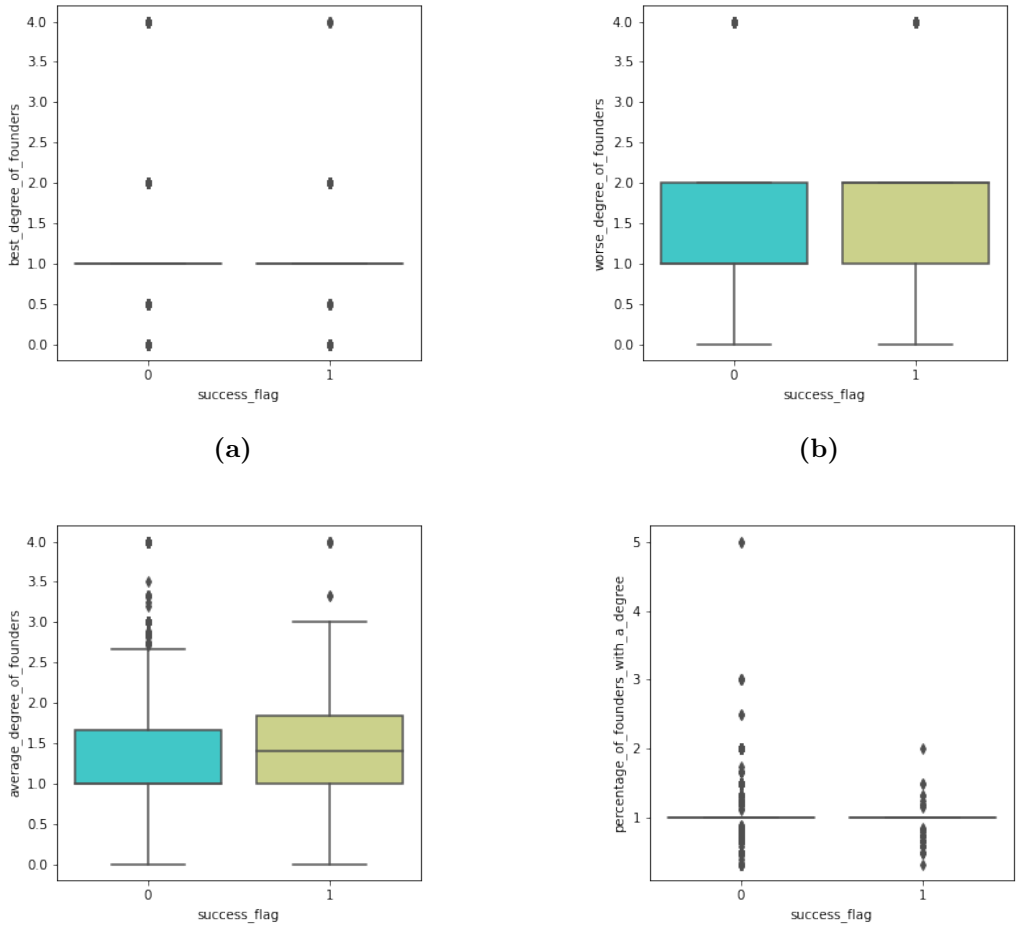
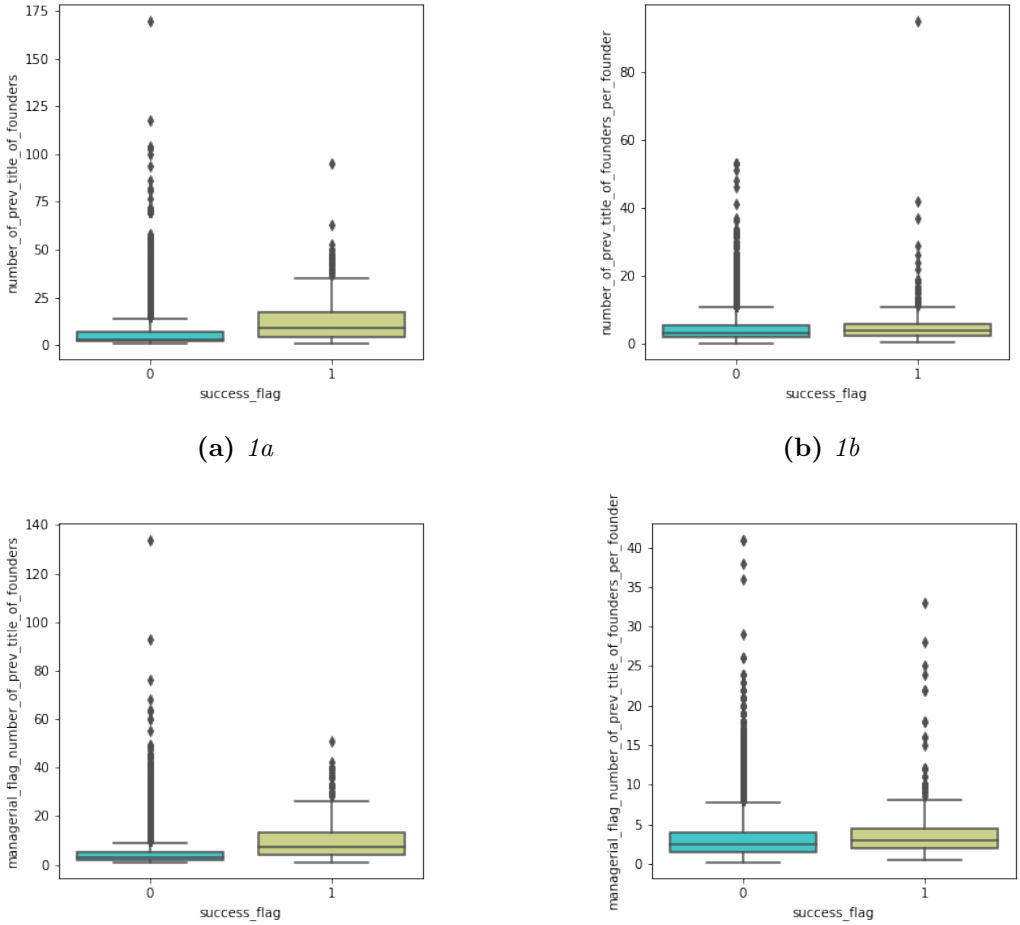


Figure 9: ***(a)*** *shows box plot distribution of success or failure of a company depending on the best degree level of the founders.* ***(b)*** *shows box plot distribution of success or failure of a company depending on the worse degree level of the founders.* ***(c)*** *shows box plot distribution of success or failure of a company depending on the average degree level of the founders.* ***(d)*** *shows box plot distribution of success or failure of a company depending on the percentage of founders that have a degree.*

Feature	Highest Education Level	Lowest Education Level	Average Education Level	Percentage of Founders Educated
Correlation	−0.0119	0.0900	0.0470	−0.0171

Table 7: *Table showing the correlation between the features and the success of the start-up*

A.5 Team Experience



(a) *1a* **(b)** *1b*

(c) *1c* **(d)** *1d*

Figure 10: ***(a)*** *shows box plot distribution of success or failure of a company depending on the experience of founders.* ***(b)*** *shows box plot distribution of success or failure of a company depending on the average experience founders.* ***(c)*** *shows box plot distribution of success or failure of a company depending on the managerial founders.* ***(d)*** *shows box plot distribution of success or failure of a company depending on average managerial experience of founders*

Feature	Number of Previous Title	Number of Previous Title Per Founders	Number of Previous Managerial Titles	Number of Managerial Titles Per Founder
Correlation	−0.0119	0.0900	0.0470	−0.0171

Table 8: *Table showing the correlation between the features and the success of the start-up*

A.6 Connectedness of Founders

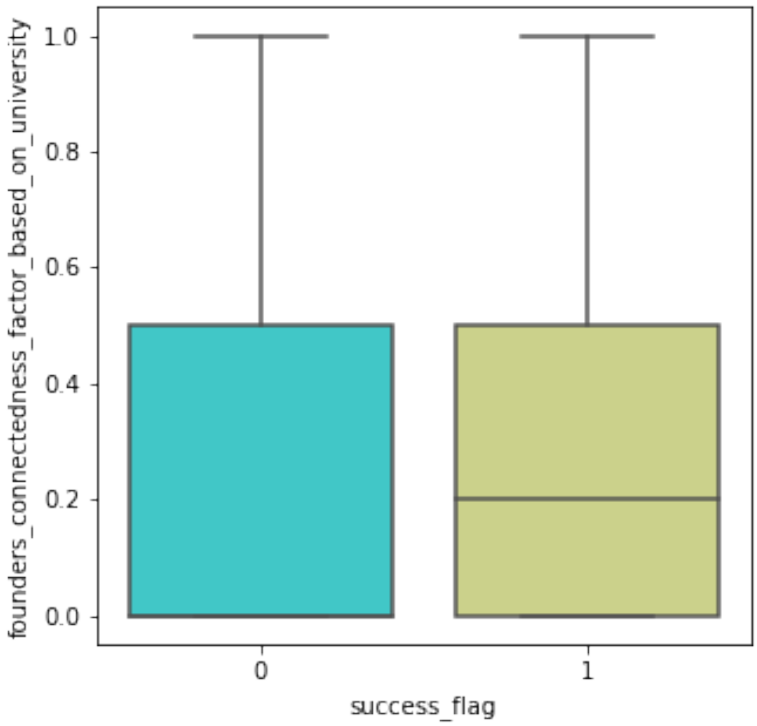
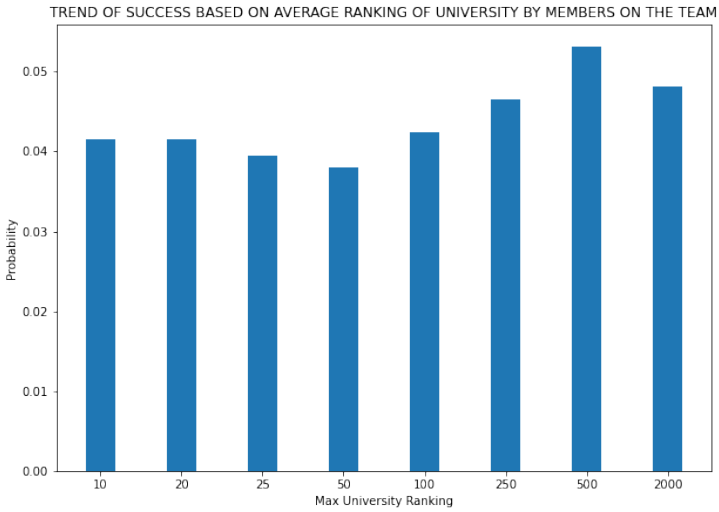


Figure 11: *Figure showing the box plot distribution based on the connectedness of the founders*

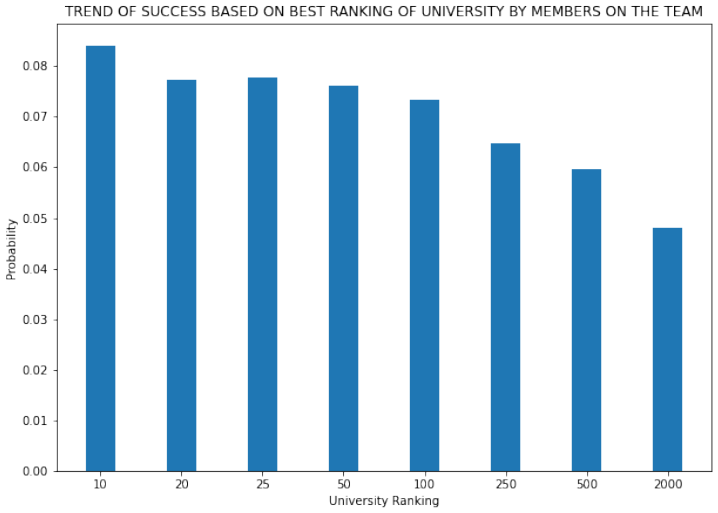
Feature	Connectedness Factor
Correlation	0.0421

Table 9: *Table showing the correlation between the features and the success of the start-up*

A.7 Trend of Success based on University Ranking



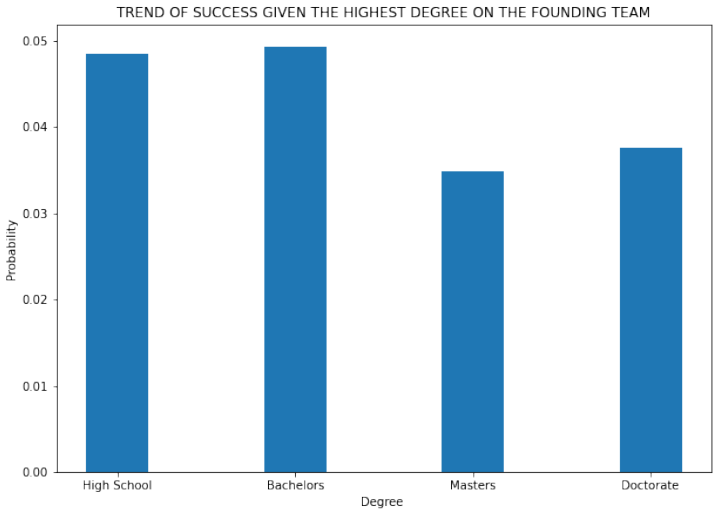
(a) *1a*



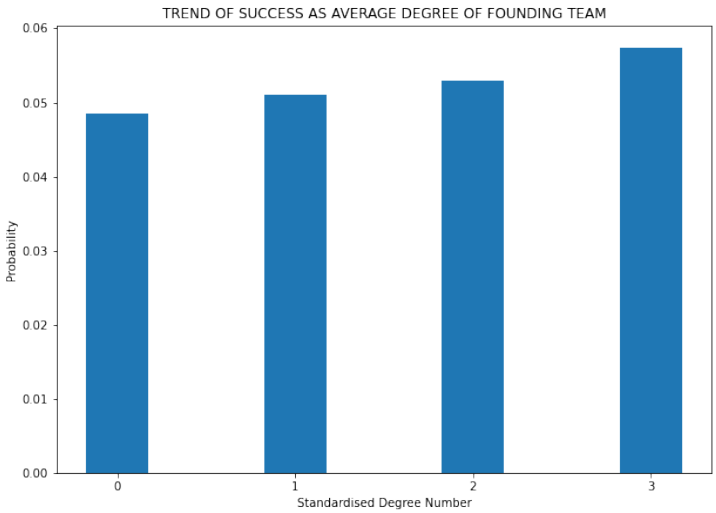
(b) *1b*

Figure 12: ***(a)*** *shows the probability of success based on having an average university ranking of at most x.* ***(b)*** *shows the probability of success based on having a best university ranking of at most x amongst the founders.*

A.8 Trend of Success based on Education of the Founding Team



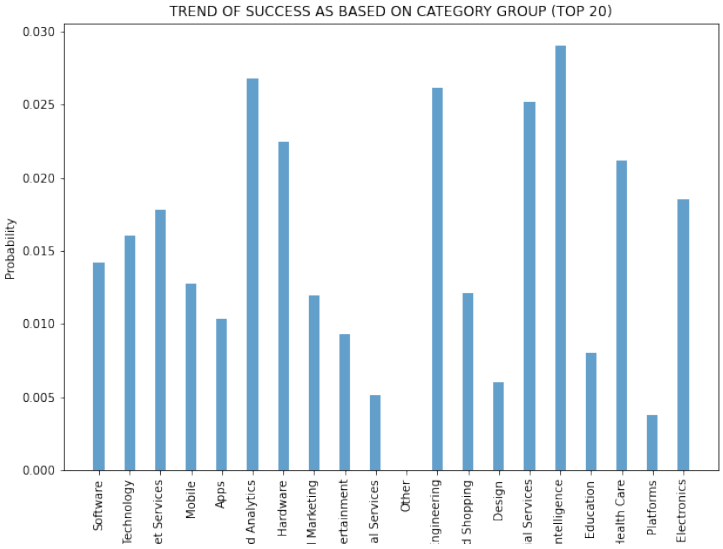
(a)



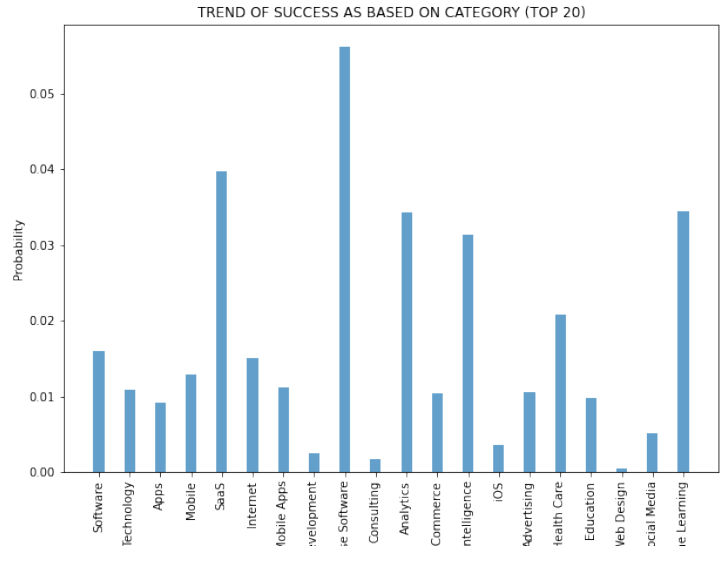
(b)

Figure 13: ***(a)*** *shows the probability of success based on having an degree level of at least x.* ***(b)*** *shows the probability of success based on having an average degree level of at least x.*

A.9 Trend of Success Based on Category and Category Group Start-up Operates In



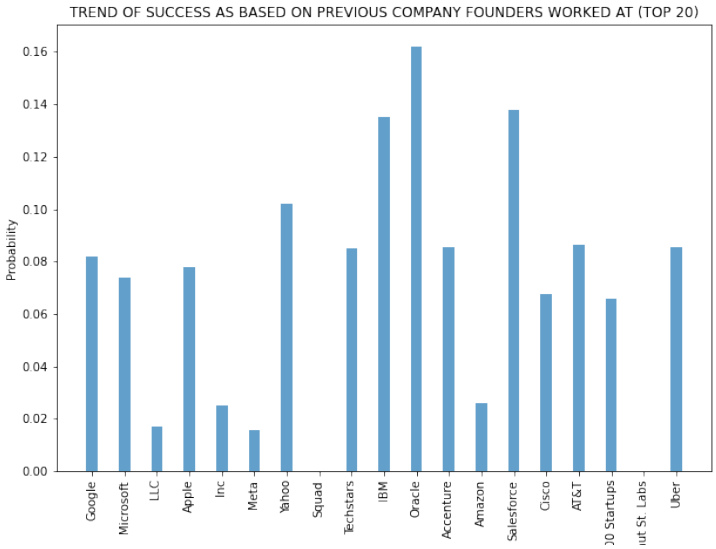
(a) *1a*



(b) *1b*

Figure 14: ***(a)*** *shows the probability of success based on the category group the company operates in.* ***(b)*** *shows the probability of success based on the category the company operates in.*

A.10 Trend of Success based on Previous Company



(a) *Figure showing the probability of success based on the previous companies the founders have worked at.*