Midway Report

Insert Subtitle Here

Khush Bawal
Computer Science
NC State University
Raleigh, NC-27606, USA
ksbawal@ncsu.edu

Samuel Burke
Computer Science
NC State University
Raleigh, NC, USA
sburke2@ncsu.edu

Sid Kaju
Computer Science
NC State University
Raleigh, NC, USA
skaju@ncsu.edu

INTRODUCTION AND BACKGROUND

Problem Statement

Navigating the Startup Landscape
In the realm of entrepreneurship, startups hold a central position, embodying innovation and ambition. They are vital to the economy and the advancement of a nation. From established tech hubs like Silicon Valley to emerging centers of innovation worldwide, these new ventures encapsulate the hopes and visions of their founders. Their mission is to disrupt industries, tackle pressing issues, and leave a lasting mark on the business landscape. However, amidst the stories of triumphant success, there are countless others that stumble and ultimately fade into obscurity. This sharp contrast between success and failure highlights the importance of understanding the intricate factors that influence the journey of startups.

Startups make up almost a whole different industry compared to traditional companies. The community of startups run very differently. Important decisions are made daily and these can make or break the company.

Predicting Startup Success: A Multifaceted Challenge
At the core of our investigation lies the mystery of startup achievement – a complex challenge that cannot be easily defined. Success for startups goes beyond just making money; it includes factors like making a significant impact on the market, being able to grow rapidly, introducing innovative ideas, and most importantly, being able to continue and flourish in a competitive environment. Conversely, failure can take many shapes, ranging from running out of money to making wrong decisions strategically, or simply becoming irrelevant in the market.

Although the product/service is at the heart of a startup, a good product does not guarantee success. Entrepreneurs often need to learn how to market, manage money, acquire customers, and much more. Startups that seem to be on the right trajectory could fall at any moment, and conversely, quick changes could help bring a company to life.

Startups do not have a fixed structure, and small details can lead to one startup being a success. Conversely, details that seem rather unimportant can lead to a startup being closed early.

Implications for Investors
As an investor, the volatile nature of startups makes it difficult to judge a startup based on surface level factors. Currently, investors usually make a decision on whether to invest based on the product/service itself, the qualities of the entrepreneur, or the numbers(sales, customer retention, etc.). By nature, investing is an activity that people get very passionate about, since they are risking their own money in hopes for a greater return. Oftentimes, internal bias and emotions play a role in investing, which can lead to risky plays. Using a data driven model will eliminate the emotional aspect of investing.

For investors, the stakes are equally high, if not higher. As the lifeblood of the startup ecosystem, investors play a pivotal role in fueling innovation and driving economic growth. Yet, their decisions are fraught with uncertainty, as they navigate a landscape rife with risk, volatility, and unpredictability. Identifying promising startups amidst a sea of contenders requires more than just financial acumen; it demands a nuanced understanding of market dynamics, industry trends, and the intangible qualities that differentiate winners from also-rans.

Implications for Founders
Founders can use the model to help optimize their success. They can learn what details will increase their chances of success(even if it is just a little bit) outside of the quality of the product/service itself. The startup space can be brutal and entrepreneurs need every advantage they can get.

The Role of Predictive Analytics
In this setting, predictive analytics becomes a crucial instrument for making informed decisions. By utilizing the extensive data produced by startups – including financial figures, operational details, market perceptions, and geographic information – predictive models provide insights that surpass traditional knowledge. These models can pinpoint early signs of success or failure, empowering investors to allocate resources strategically and manage risks efficiently.

In conclusion, the quest to understand the determinants of startup success is a multifaceted endeavor with far-reaching implications for founders, investors, and the broader entrepreneurial ecosystem. By delving into the intricacies of this enigma, we seek to unravel the underlying patterns and dynamics that shape the fate of startups. Through the lens of predictive analytics, we aim to

illuminate the path forward, empowering investors with the insights they need to navigate uncertainty and maximize returns in the pursuit of innovation and growth.

Related Work

1. Robert N. Lussier & Claudia E. Halabi (2010) A Three‐Country Comparison of the Business Success versus Failure Prediction Model, Journal of Small Business Management, 48:3, 360-377, DOI: 10.1111/j.1540-627X.2010.00298.x
    - Views companies in context of the economy and nature of their respective countries. The different countries are used as a comparison of economic and location data.

2. A Business Success Versus Failure Prediction Model for Entrepreneurs with 0-10 Employees. (1996). Journal of Small Business Strategy (archive Only), 7(1), 21-36. https://libjournals.mtsu.edu/index.php/jsbs/article/view/327
    - Uses generic non financial data and survey results from experts to find out the predictors of success

3. Why do startups fail? A core competency deficit model. Front. Psychol., 07 February 2024 https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1299135/full
    - Focuses more on the reasons why startups fail

4. Why Do Startups Fail And Can We Create An AI Formula To Prevent Failure? An Interview With Harvard Business School Professor Tom Eisenmann. (Taarini Kaur Dang) Why Do Startups Fail And Can We Create An AI Formula ...Forbeshttps://www.forbes.com › Innovation › Big Data
    - Explores how to use AI to predict startups

5. The 3 biggest reasons startups failed in 2022, according to a poll of almost 500 founders (CNBC) https://www.cnbc.com/2023/01/20/top-reasons-why-startups-failed-in-2022-study.html

METHODS

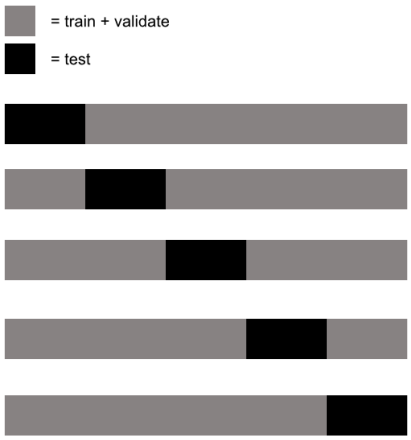
Novel Aspect

Approach

We have formed our question into the following classification question: is a given company X likely to be successful given the parameters in our classifier, $P(Y \neq closed|x) > P(Y = closed|x)$. A company is assumed to be successful if the company remains open for at least ten years or is acquired by another company after any duration. We decided to use the Logistic Regression model for our classification problem. We used linear dependencies for each attribute that we deemed useful in the dataset and added binary attribute columns for categorical data. To maximize the likelihood of a binary result, we used the typical linear regression to predict the log-odds of having class 1(successful) vs 0(unsuccessful). The training objective uses the cross-entropy loss function to maximize the likelihood of a binary result. To minimize loss we minimize the negative log of the predictor $P(Y = Y_i|x)$:

$$-log(P(Y = Y_i|x)) = log(P(Y = 1|x)) - (1 - Y_i)log(P(Y = 0|x))$$

For testing, we split the data into 80% training data and 20% testing data. Furthermore, we perform k-cross fold validation, to give a more accurate picture of the overall performance of our model, with k = 5 folds.



Rationale

The rationale for selecting Logistic Regression is that the model will be able to be trained fast, and generate results with low computational effort. Furthermore, using Logistic Regression will help us understand to a greater extent what attributes have a large impact on making a company successful or unsuccessful by looking at the weights of each parameter. Some disadvantages to using Logistic Regression in this study is that we assume there are linear boundaries, which may or may not capture the relationships best. However, we believe the attributes such as what city a company is started and the field the company works in are significant factors in determining if the company is successful and will perform well with a Logistic Regression model. We use cross validation to assess the overall performance of the model, instead of the risk of testing an overfitted version.

PLAN & EXPERIMENT

Dataset

Our original dataset from Kaggle about the Success predictions of Startups. The dataset contains several different attributes that describe startups. The attributes cover a wide variety of details such as the location, founding date, information regarding the funding rounds, the industry of the business, and the time in between funding rounds. These attributes will help us assemble the properties that define each company. There are also attributes to indicate whether the startup was a success or not. One attribute tells us whether the company is a top 500 company and another tells us if the company was shut down or acquired. There are also a ton of attributes that are highly specific that could have a big impact on a company. For example, there is a column that indicates the location of the startup for each of the states California, Texas, NY, and MA, probably because these states have a much higher impact than others. There are similar columns to add industry tags to each startup. The location data is also extensive, there are attributes telling us the exact latitude and longitude, zip code, city, and more. This allows us to be flexible with how we want to weigh sub locations differently. The dataset has 924 data points, giving us a large sample size of companies.

Our Dataset had a lot of noise and outliers which hindered the feature selection process. Some of the discrepancies were:-

1. There were repeat analytical columns which we dropped before proceeding to the next steps like 'Zip Codes' as we were already relying on 'longitude', 'latitude', 'cities' and 'states' to make analysis based on location.
2. We even dropped a few undocumented columns which have no reasoning like 'Unnamed: 0', 'id', 'Unnamed: 6', 'labels', 'object_id', and 'is_othercategory'. All of these features had no direct relation with the outcome while some did have some influence over the overall outcome; they were too insignificant to be considered for our model.
3. Other than the feature selections or filtration we also had to do a lot of searching to clean the data for example in Column G 'city' many companies have either misspelled or wrote each city name differently, like 'Vienna - Viena' or 'New York - NY - New York City' or 'El Sugundo - El Sugundo,' or 'Sunnnyvale - Sunnyvale'. These minor details after correction were then imported into data frames if we hadn't done these steps before manually coding out some of these parts would have been tedious. However, we did leave more of the feature selection parts to the model.

The above described were the parts where we clean and prepare the data to further enhance the features to give us parameters that are easily accessible. For example, we use the 'founded_at' and 'closed_at' (this includes closed after acquired) to give us 'Operational_Span' which indicates the difference between both. This feature would be evaluated in months and the companies who are still operational will have the parameter of -1 which will be further used in the evaluations in the future. Luckily some parts of our dataset have already been converted to binary data seen from the columns 'is_CA', 'is_NY', 'is_MA', and 'is_TX' while others are already in binary from columns AD - AV. This makes it easier for us to implement our Logistic regression model.

Some additional decisions we took was to limit the range of these companies from global to just the USA, so basically we will be limited our latitude and longitude axis values to the borders of 'Maine', 'Florida', 'Washington', 'Alaska', 'California'.

These are currently the few changes we made to our attributes and values, some were made manually while most were made using the pandas library in Jupyter Notebook.

Hypothesis

We are predicting that situational factors such as location, network (helps with funding), money, and the category will matter just as much or more than the quality and innovation of the product/service itself.

Experimental Design:

Libraries we are using :-

1. Pandas - we used to convert files and then convert it to dataframes to use them for training and testing.
2. Numpy - for array based calculations.
3. Sklearn - for some preprocessing, training, testing and applying our logistical regression.
4. Matpotlib - we are using this to add graphs and visualization to help us graph the concepts of logistic regression.
5. Seaborn - for heat maps and more visualization tools.

RESULTS

Currently, we are assembling results. We plan on using the Cross-Validation method to test our model, with the dataset split up into 5 sections. We will then take the mean of the 5 results to calculate the overall result. We plan on computing a variety of metrics, such as accuracy, recall, precision, and F1 score. We hope to get a high rate on all of these metrics.

We are also going to visualize (using the libraries matplotlib, folium, and autowiz) our data. We plan on showing a map with the locations of all the startups in the dataset shown. We will also use charts and graphs to compare the results against attributes that are commonly considered huge for startup success.

Using the results, we will also try to find some relationships between the features and the success of a company, as well as the relationships between certain features. We will use this to make some inferences, backed by the results of our model.

The results and the impact of each one will be explained in detail, and our inferences from these will be explained and compared to current work on the subject. From this, readers will be able to form their own opinions.

Broader Impacts

Startups are the bases for innovation and some of the main reasons for these startups fail/succeed depends upon how the market reacts to their product. Either they survive through the competitive consumer market and get acquired by one of the top MNCs like ‘Microsoft’, ‘Alphabet’, etc. Or they never get acquired and shut down before making an impression or impact on their consumers. Our softwares helps to assess these failures and learn from both the success and failure scenarios to help ongoing startups predict their future path.

One reason why we encourage startups to succeed is to take down monopolies in the industry, these giants control the cash flow and the power in the entire country and have a huge influence on the selling point of all the products in the market hence it is very important for us to make sure the fresher have a chance to have some control over the market and in this process innovation is never restricted or capitalized by the unethical groups.

MEETING ATTENDANCE

- 4/3/24 Khush Bawal, Samuel Burke - Preliminary ideas and concepts regarding the project.
- 4/9/24 Khush Bawal, Samuel Burke, Sid Kaju - selecting the approach and design of the model.