Startup Success Prediction

1st Chinmay Patankar
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
cpatanka@stevens.edu

2nd Mandar Parab
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
mparab1@stevens.edu

3rd Kush Jani
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
kjani1@stevens.edu

***Abstract*—Startups are companies or ventures that are focused on a single product or service that the founders want to bring to market. Startups play a major role in economic growth and a few are valued at over US 1 billion dollars. So how do we know whether a startup is successful or not? The success of a startup is defined based on the amount of money they acquired through the process of Merger and Acquisition or an IPO (Initial Public Offering). To solve this we are using a data set that contains detailed information of a startup including its various stages of funding to identify characteristics of a successful startup. Identifying these characteristics is one of the main aims of the project. Creating a visual representation of the features helps us identify which are significant. For this we are using three machine learning algorithms which are Logistic Regression, K Nearest Neighbour (KNN), and Random Forest.**

I. Introduction

Our main purpose is to find out whether a presently operational startup will succeed or fail. We use a startup success prediction data set for this, which includes both quantitative and categorical factors that aid in making a solid prediction. We seek to achieve this goal by combining Machine Learning techniques with the Python programming language. We can use machine learning algorithms to predict if a startup will succeed or not. The goal of this project is to provide a variety of insights based on classification. We will utilize three algorithms for this project: Logistic Regression, K Nearest Neighbour (KNN), and Random Forest.

II. Related Work

A group of scientists(Phys.Org) was able to develop an AI that could accurately predict whether a startup will end up being the next big thing, or it will be another failed attempt to disrupt the market. The machine learning models that were used for this AI predicting tool look into over a million companies to further say the scalability of a startup. That said, if the tool turns out to be absolutely accurate, more investors could help the capitalization of these budding startups. On top of that, it would also be beneficial to investors as they will no longer be risking their money for startups that would end up being a waste of their time.

III. Our Solution

With the help of machine learning algorithms to predict a success-rate of a startup with a variety of insights based on the features such as agefirstfundingyear, agelastfundingyear,agefirstmilestoneyear, agelastmilestoneyear, relationships, fundingrounds, fundingtotalusd, milestones, categorycode, hasV, hasangel, hasroundA, hasroundB, hasroundC, hasroundD, avgparticipants, istop500 our goal is to predict whether a particular startup is successful or not and identify key features that contribute to a startup's success.

A. Description of Dataset

The dataset is not large. It has 48 columns/features and 924 rows. The data contains industry trends, investment insights and individual company information about a startup like name, state(which state the startup is located in), city(city in which the startup has its main office), categorycode (which industry does the startup belong to), it also includes individual industry features such as issoftware, isweb, ismobile, isenterprise, isadvertising, isgamesvideo, isecommerce, isbiotech, isconsulting, zipcode, latitude (Latitude coordinate for Global Coordinate System,), longitude (Longitude coordinate for Global Coordinate System), fundingtotalusd (total amount of vending received), foundedat (year the startup was found in), closedate (year the startup was shutdown), agefirstfundingyear (age of the startup when it received its first funding), agelastfundingyear (age of the startup when it received its last funding), firstfundingat (year the startup received its first funding), lastfundingat (year the startup received its last funding), relationships, fundingrounds (number of funding rounds the startup went through), milestones (number of milestones in the startup's lifespan), agefirstmilestoneyear (age of the startup at its first milestone), agelastmilestoneyear (age of the startup at its last milestone), hasVC (whether startup received funding from Venture capital or not), hasangel (whether startup received angel funding or not), hasroundA (whether startup received round A funding or not), hasroundB (whether startup received round B funding or not), hasroundC (whether startup received round C funding or not), hasroundD (whether startup received round D funding or not), avgparticipants (average number of funding received), istop500 (startup is top 500 or not), status(acquired/closed - the target variable, if a startup is 'acquired' by some other organization, means the startup succeed).

Unnamed: 6	name	labels	...	object_id	has_VC	has_angel	has_roundA	has_roundB	has_roundC	has_roundD	avg_participants	is_top500	status
NaN	Bandsintown	1	...	c:6669	0	1	0	0	0	0	1.0000	0	1
NaN	TriCipher	1	...	c:16283	1	0	0	1	1	1	4.7500	1	1
San Diego CA 92121	Plixi	1	...	c:65620	0	0	1	0	0	0	4.0000	1	1
Cupertino CA 95014	Solidcore Systems	1	...	c:42668	0	0	0	1	1	1	3.3333	1	1
San Francisco CA 94105	Inhale Digital	0	...	c:65806	1	1	0	0	0	0	1.0000	1	0
Mountain View CA 94043	Matisse Networks	0	...	c:22898	0	0	0	1	0	0	3.0000	1	0
NaN	RingCube Technologies	1	...	c:16191	1	0	1	1	0	0	1.6667	1	1
NaN	ClairMail	1	...	c:5192	0	0	1	1	0	1	3.5000	1	1
Williamstown MA 1267	VoodooVox	1	...	c:1043	1	0	1	0	0	1	4.0000	1	1
NaN	Doostang	1	...	c:498	1	1	1	0	0	0	1.0000	1	1

Fig 1. Sample of the Dataset

The source of this dataset is from kaggle. As this dataset has a lot of features we have therefore dropped unnecessary and redundant features. There were some null value entries within the dataset which we removed and cleaned from the main dataset.

B. Machine Learning Algorithms

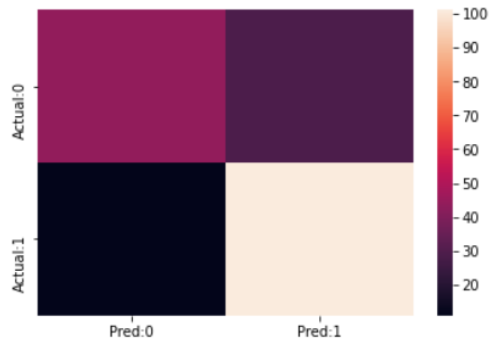
We are using following three algorithms:

1. Logistic Regression

2. K Nearest Neighbor (KNN)

3. Random Forest

Currently we have implemented Logistic Regression Algorithm. The Logistic Regression algorithm was chosen as it works best when there is a binary output. Our dataset has two target classes 'acquired' and 'not acquired' which we have encoded with 1 and 0 in the status column. The accuracy from Logistic Regression we got is 78 percent with percent of acquired correctly predicted is 90 percent and not acquired is 60 percent as shown in Fig 2.




```
          Pred:0  Pred:1
Actual:0      44      29
Actual:1      11     101
Percent Acquired correctly predicted:  90.17857142857143
Percent Not Acquired correctly predicted:  60.273972602739725
Accuracy of the test set:  0.7837837837837838
```

Fig 2. Accuracy rate of Logistic Regression

K Nearest Neighbor (KNN) algorithm is a supervised algorithm. Based on the most relevant features we will be using KNN to predict whether a startup is successful or not. This will be working according to the nearest neighbor.

Random Forest classification algorithm will be used. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. By developing multiple decision trees and considering the majority of the output values from these decision trees we will be able to correctly predict whether a startup is successful or not.

C. Implementation Details

Project includes, applying algorithms such as Logistic Regression, KNN and Random Forest to obtain insights and make predictions. First comes the prepossessing steps, where all the missing values were filled with zero and irrelevant data were removed. Missing values were causing incorrect visualization of some features. Considering this problem, all the missing or null values were removed successfully. There were some redundant columns like issoftware, isweb, ismobile, isenterprise, isadvertising, isgamesvideo, etc. which were combined into one single column called category codes. Out of 48 columns, 31 columns/features were dropped from final data to be used further.

funding_total_usd	milestones	category_code	has_VC	has_angel	has_roundA	has_roundB	has_roundC	has_roundD	avg_participants	is_top500	status
3352194	2	analytics	0	1	1	0	0	0	9.50	0	0
1200000	3	travel	0	1	0	0	0	0	5.00	1	1
11040000	2	software	0	0	1	1	1	0	2.50	1	1
13400000	1	public_relations	0	0	0	1	0	0	2.00	1	1
20000000	1	security	1	0	0	0	1	0	2.00	1	1
15200000	0	software	0	0	1	1	0	0	2.00	1	1
18500000	3	mobile	0	1	1	1	0	0	2.25	1	1
5500000	2	web	0	0	1	0	0	0	2.00	1	1
30000000	1	biotech	0	0	0	0	1	0	9.00	1	1
1500000	3	fashion	0	0	0	0	0	0	1.00	0	1

Fig 3. Sample of the Dataset after Data cleaning and Data pre-processing

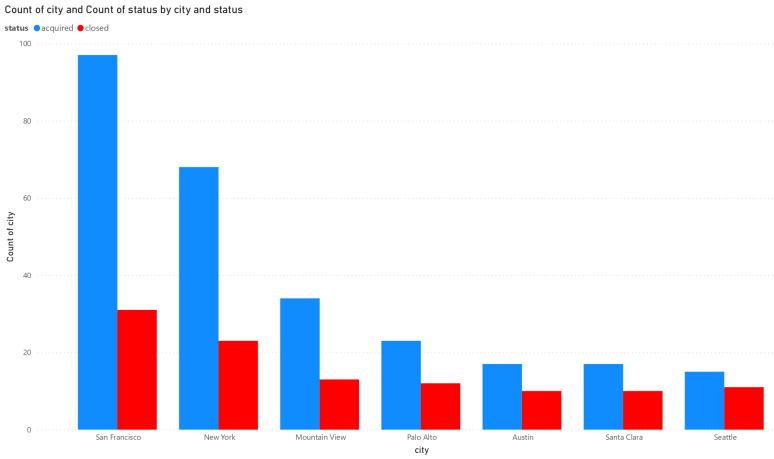


Fig 4. Which City has the most number of successful startups

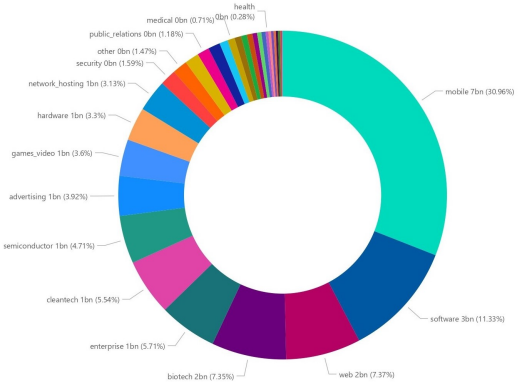


Fig 5. Which industry has received largest amount of funding

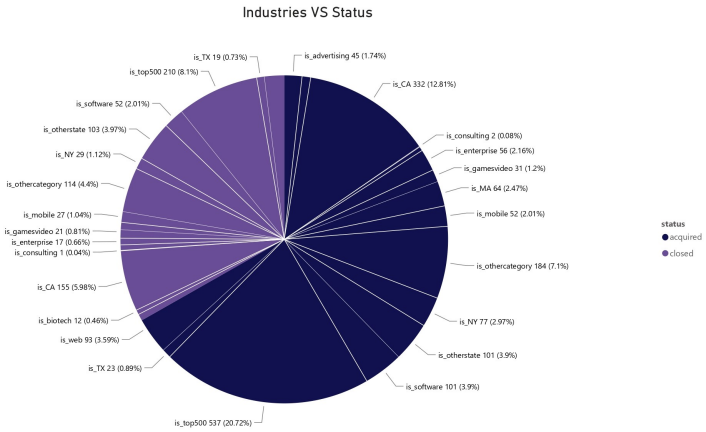


Fig 6. Which industry has the most number of successful startups

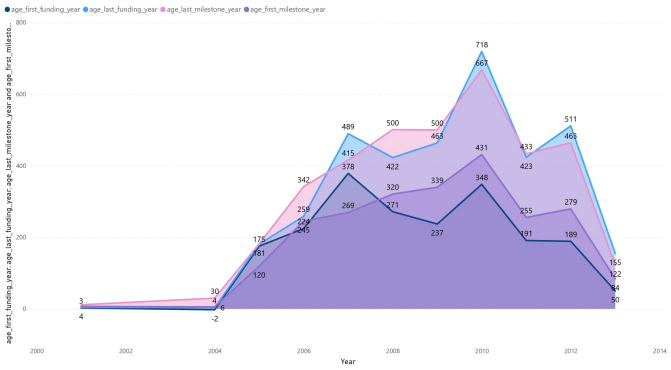


Fig 7. Funding received over the years including milestone years

We have made different visualizations for better understanding of the data. In Fig 4, we have shown which city in the United States of America has the highest number of successful and unsuccessful startup in which San Francisco holds the highest rank. An important insight from this data was the amount of funding received by a startup of a particular industry. In Fig 5, this understanding is visually represented using a pie chart. Related to Fig 5, the pie chart in Fig 6, describes which industry has the most number of successful and unsuccessful startups. In Fig 7, we visually represent amount of funding received starting from its funding round to the last funding round including a startup's milestone years. These representation help us better understand the data and helps us choose important features. Following are the algorithms which we are going to use in the project.

Logistic Regression (also called Logit Regression) is commonly used to estimate the probability that an instance belongs to a particular class.If the estimated probability is greater than 50 percent, then the model predicts that the instance belongs to that class (called the positive class, labeled "1"), or else it predicts that it does not (i.e., it belongs to the negative class, labeled "0"). This makes it a binary classifier.It uses a logistic function also called the sigmoid function(sigmoid function = 1 / (1 + e^{-value})that outputs a number between 0 and 1.

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. KNN performs better with a lower number of features than a large number of features.

Classification in random forests employs an ensemble methodology to attain the outcome. The training data is fed to train various decision trees. This dataset consists of observations and features that will be selected randomly during the splitting of nodes. A rain forest system relies on various decision trees. Every decision tree consists of decision nodes, leaf nodes, and a root node. The leaf node of each tree is the final output produced by that specific decision tree. The selection of the final output follows the majority-voting system. In this case, the output chosen by the majority of the decision trees becomes the final output of the rain forest system. The diagram below shows a simple random forest classifier.

IV. Comparison

Currently implementation of Logistic Regression is completed with good accuracy as there is no other algorithm implemented at this point of time. So we cannot compare the accuracy but this can be done after the implementation of other algorithms.

V. Future Directions

In the future we would like to identify common patterns or traits between different successful startups and failed startups and predict whether a startup would be successful or not.We would find the most accurate method for prediction. We will need to clean the data more to find clear relationships between the data and features. Provided some more time, we could increase the accuracy of correctly predicting a successful startup.

VI. Conclusion

Currently, we are in the process of implementing our Machine Learning algorithms. We currently have Logistic Regression implemented and we will compare that with other algorithms to determine a more accurate algorithm so we are unable to provide a comprehensive conclusion.

References

https://www.techtimes.com/articles/265116/20210908/this-ai-could-predict-startup-success-research-claims-it-has-90-accuracy.htm

https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3878-2019.pdf