PREDICTING STARTUP SUCCESS

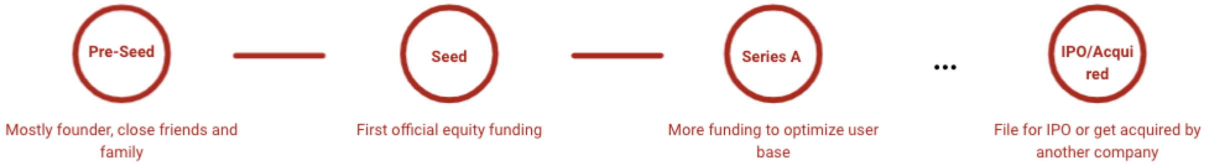
08/04/2021

By Prianka Ball, Kalyani Pavulur, Shivprasad Chavan
pball@saintpeters.edu , kpavuluri@saintpeters.edu, schavan@saintpeters.edu

INTRODUCTION

Startup refers to a company that is in their first stages of operations. Startups can be founded by one or more entrepreneurs who are working on developing a product or service. Startups normally have very high cost and limited revenue. As they require a lot of capital to take it off the ground, they look for capital from a lot of sources like venture capitalists.

Startups go through multiple rounds of funding to raise capital. The different funding rounds that let outside investors the opportunity to invest cash in exchange of equity or partial ownership of the company. Other types of investments are debt, convertible note, stock or dividends. Startups can start off with “seed” funding or angel investor funding at the beginning. The next funding rounds can be followed by Series A, B, C and so on. Goal of most startups is to get acquired by a different company or become a publicly traded company.


Pre-Seed
Mostly founder, close friends and family
Seed
First official equity funding
Series A
More funding to optimize user base
...
IPO/Acquired
File for IPO or get acquired by another company

90% of startups fail due to bad product market fit, marketing problems, team problems or other issues. They also fail within the first few years. This makes startup investment very risky. Historically only venture capitalists could invest in startups but due to the recent trend in crowdfunding sites, an average investor can easily grab a piece of an exciting startup.

Problem Statement

Startup investment can be very risky due to the high failure rate of startups. People like angel investors and venture capitalists have a very high risk while they are investing in startups.

To assist startup investors with their decisions, in this project we aim to find the important features that lead to startup success and forecast a company’s success with supervised machine learning methods.

DATA

To train the machine learning model, we used investment data about startup companies available on Kaggle. The data has been collected from Crunchbase which is a leading website for company insights from early stage startups to Fortune 1000.

The data had around 54k rows and 39 columns. The dataset had company information such as name of the company, url, market, country, state, region, city, founded date, first funding date, last funding date. It also had data on different investment types such as seed, venture equity crowdfunding, undisclosed funding, convertible note, debt financing, angel, grant, private equity, post ipo equity, post ipo debt, secondary market, product crowdfunding, round A-H series funding. Detailed descriptions of the different funding types is available here. Status of the companies were also available and segmented by acquired, operating and closed.

METHODOLOGY

Before we could use the data to train the different models, we had to clean the data and select the most important columns to be included into the model. One of the biggest problems we had with the dataset was that it had a lot of zeros and a lot of columns to choose from.

We also realized later that the status column had around 80% of the companies as operating status and the rest as closed and acquired companies.

Data Cleansing Methodology

To clean the data, we removed extra spaces from different columns and also removed things like " , ", " - " where ever necessary. We made sure that columns that had numbers were being read as numbers and also converted all of the date columns into date data types.

Rows with null values were removed. Columns with a high percentage of null values like state, city, region, and found date were removed.

Feature Engineering

We created a new column that included differences in time between the last and first funding date.

A total investment column was also created that included the sum of all the investments (seed, venture, equity crowdfunding, undisclosed, convertible note, debt financing, angel, grant, private equity, post ipo equity, post ipo debt, secondary market, product crowdfunding).

The data contained 753 different market values. As there were a lot of different types of markets, we wanted to reduce this number. To reduce the number of markets, we grouped markets into different industry groups segment on the industry grouping list produced by crunchbase. The list can be found here. The new column Industry Group had 43 industry groups .

The data contained 115 countries. The dataset was joined with a different dataset that contained the country name and continent. We made a new column for the continent name.

Numerical values like total investment, difference in funding year, funding rounds, seed, venture were turned into categories like low and high based on their spread in number. The categorical values were again turned into numbers for the models to understand.

Other columns like equity crowdfunding, undisclosed, convertible note, debt financing, angel, private equity, post ipo equity, secondary market, product crowdfunding, round A -H were turned into 0 and 1 based on if the company was able to raise that type of funding. This was done as the columns had too many zero values.

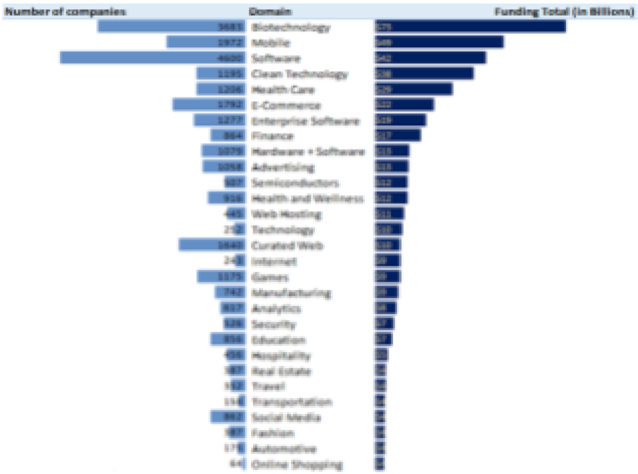
ANALYSIS

The analysis of the dataset was important to understand what is important to use in the final model. It also helped us understand the data more before using them in the models.

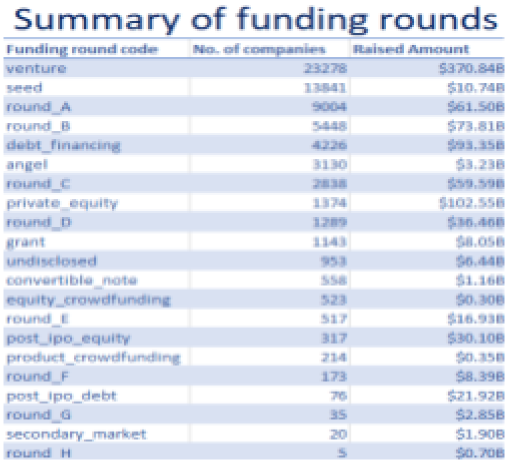
Exploratory Data Analysis

The analysis below contains some of the important EDA we did on the data.

Most of the companies were in the Software and Biotechnology industry. Biotechnology had the highest number in total funding. Mobile companies had the second lowest number in total funding.



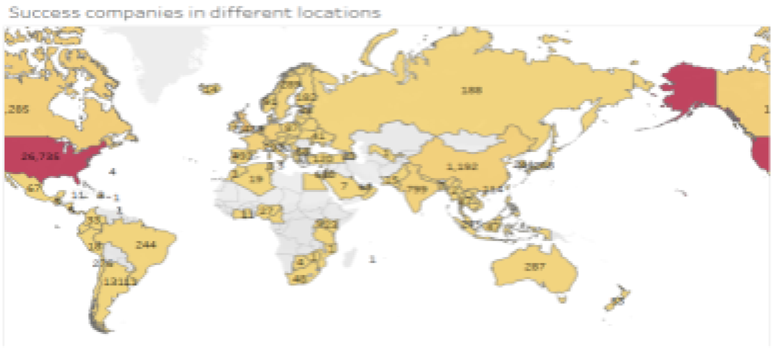
Total funding in different type of companies



Summary of funding rounds

Funding round code	No. of companies	Raised Amount
venture	23278	$370.84B
seed	13841	$10.74B
round_A	9004	$61.50B
round_B	5448	$73.81B
debt_financing	4226	$93.35B
angel	3130	$3.23B
round_C	2838	$59.59B
private_equity	1374	$102.55B
round_D	1289	$36.46B
grant	1143	$8.05B
undisclosed	953	$6.44B
convertible_note	558	$1.16B
equity_crowdfunding	523	$0.30B
round_E	517	$16.93B
post_ipo_equity	317	$30.10B
product_crowdfunding	214	$0.35B
round_F	173	$8.39B
post_ipo_debt	76	$21.92B
round_G	35	$2.85B
secondary_market	20	$1.90B
round_H	5	$0.70B

A lot of the companies raised venture and seed funding. The number of companies decreased as the companies proceeded to more series funding. Round G and H have a very low number of companies compared to round A and round B.



Places where acquired and operating companies are based at

Most of the acquired and operating companies are from the U.S

Acquired companies had higher mean and median funding compared to closed and operating companies.

status	count	mean	std	min	25%	50%	75%	max
acquired	3692.0	2.075578e+07	1.079477e+08	0.0	1100000.0	6000000.0	19500000.0	5.700000e+09
closed	2603.0	7.023194e+06	3.888355e+07	0.0	30000.0	500000.0	4000000.0	1.567504e+09
operating	41829.0	1.295244e+07	1.633604e+08	0.0	50000.0	999857.0	6000000.0	3.007950e+10

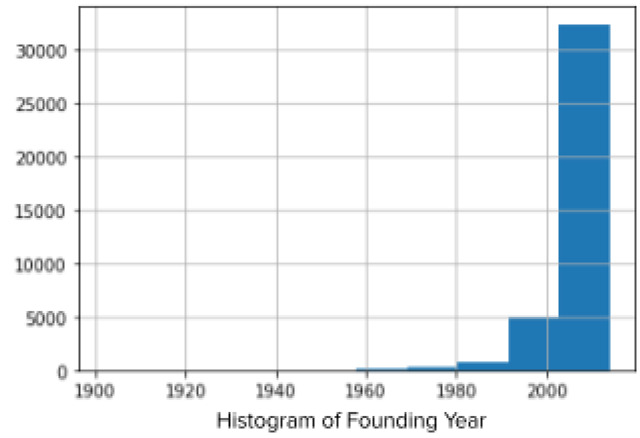
Funding total of companies with different status

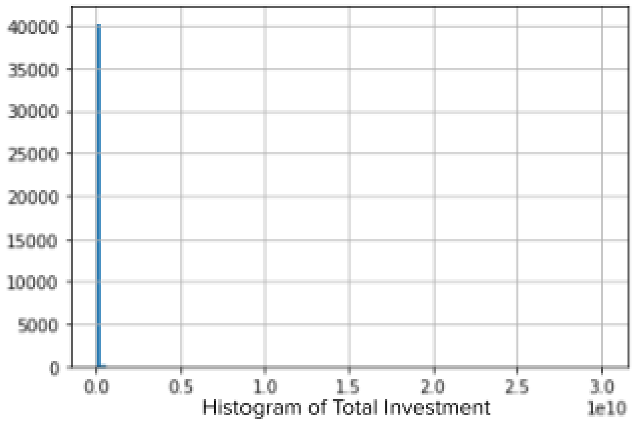
Acquired companies also had more number of funding rounds compared to companies with closed and operating statuses.

status	count	mean	std	min	25%	50%	75%	max
acquired	3692.0	2.013814	1.398832	1.0	1.0	2.0	3.0	15.0
closed	2603.0	1.434114	0.965478	1.0	1.0	1.0	2.0	11.0
operating	41829.0	1.689522	1.302072	1.0	1.0	1.0	2.0	18.0

Funding rounds of companies with different statuses

In terms of year, 2014 was the newest year and the oldest year was 1902. Most of the companies were founded quite recently around the 2000. The total investment data was very skewed, similar to the other type of funding.





Statistical Analysis

To make a decision about which columns to pick on the final model, we did a correlation matrix. The results of the correlation is below:

```
cat_status                  1.000000
cat_Industry_Group          0.027022
cat_Continent_Name          0.047636
cat_funding_rounds          0.084055
cat_diff_funding_year       0.079210
cat_total_investment        0.149947
cat_equity_crowdfunding    -0.006529
cat_venture                 0.144481
cat_seed                   -0.048227
cat_undisclosed             0.007443
cat_convertible_note       -0.003928
cat_debt_financing          0.015993
cat_angel                  -0.036605
cat_grant                  -0.008242
cat_private_equity          0.016931
cat_post_ipo_equity        -0.001666
cat_post_ipo_debt           0.000422
cat_secondary_market       -0.001582
cat_product_crowdfunding   -0.003177
cat_round_A                 0.083303
cat_round_B                 0.107367
cat_round_C                 0.104815
cat_round_D                 0.075281
cat_round_E                 0.040294
cat_round_F                 0.014688
cat_round_G                 0.002535
cat_round_H                -0.000791
```

Because of low correlation, we decided to leave out crowdfunding, undisclosed, convertible note, grant , post ipo equity, post ipo debt, secondary market, product crowdfunding, round G, round H from the final model.

MODELS

We used different types of models on the data to understand which model would be the best. We have tested with both multi class and binomial classification models. For multi class models, we were trying to predict for closed, acquired and operating companies. For binomial classification models, we tried to predict for closed and acquired companies only.

As the dataset contained 80% of the data that had operating companies, the multiclass model was good at predicting for operating companies.

For most of the models, 20% of the data was accounted for as a test dataset and results were drawn at random.

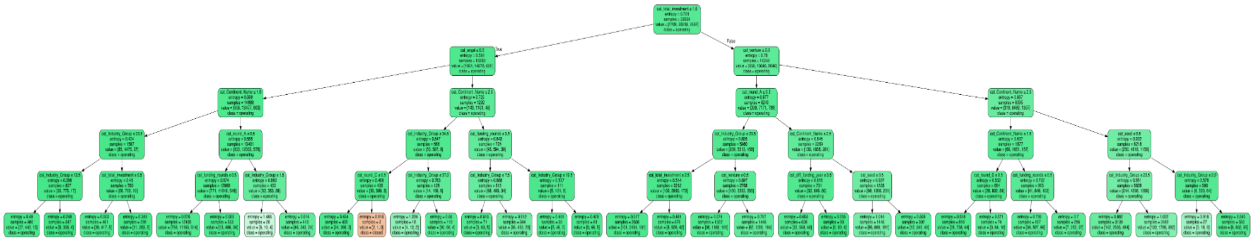
We also tested all models with undersampled data. However as biases get included into the result, we decided to leave it aside from the final analysis.

Decision Tree

Decision Tree is a flow-chart like tree structure where the internal mode is considered as a feature and the branches are considered as a decision rule. It learns to partition on the basis of the attribute value. We used decision tree on both multiclass and binomial model

Multi-classs Classification

When we decided to use only default values on the decision tree model, the model was overfitting. We used cost_complex_pruning_path from sklearn to find the most efficient alpha value that would help us prune the tree. The alpha value that would give the highest accuracy rate was used in the model. To tune the model, we also tested with grid search. Tuning with cost_complex_pruning_path and grid search gave the same accuracy results but their decision tree looked very different from each other.



Decision Tree from grid search

Using grid search, we tried to find features that would get the best accuracy rate. Our grid search gave us the best result when the criterion was entropy, max depth of 5, min sample leaf of 1 and min sample leaf of 2. According to the model total investment, continent name, and venture were the most important features.

Features	Feature Importance
cat_total_investment	0.498248
cat_Continent_Name	0.127891
cat_venture	0.121483
cat_Industry_Group	0.099878
cat_round_A	0.051733

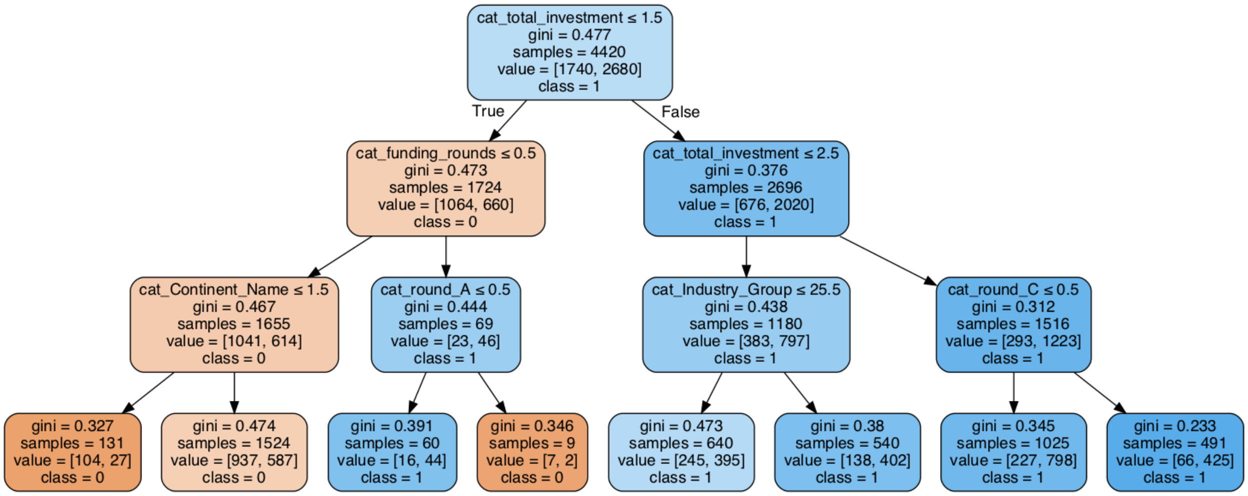
The accuracy rate came up to be 0.86.

Target Status	Precision	Recall	F1-score	Support
Closed	0.00	0.00	0.00	456
Operating	0.86	1.00	0.93	7037
Acquired	0.00	0.00	0.00	663

Classification Report of Decision Tree Multiclass Model

Binomial Classification

For the Binomial Classification, we tried to predict for closed and acquired companies. Using grid search, we tried to find the model parameters which give the best accuracy rate. We received the best accuracy rate when we used gini criterion, max depth of 3, min samples leaf of 1 and min samples split of 2.



According to this model, total investment, funding rounds, industry group, continent name were important features in understanding if a company will be successful or not. The model shows that if total investment is very important and if it is less then the company is likely to be closed.

Features	Feature Importance
cat_total_investment	0.889499
cat_funding_rounds	0.033746
cat_Industry_Group	0.027641
cat_Continent_Name	0.022541
cat_round_C	0.014658

The accurate rate came up to be 0.69 and it was good at predicting for both closed and acquired companies.

Target Status	Precision	Recall	F1-score	Support
Closed	0.60	0.56	0.58	425
Acquired	0.74	0.77	0.75	680

Classification Report of Decision Tree Binomial Model

Random forest

Random forest consists of a large number of individual decision trees that operate on ensemble. Ensemble method means that multiple models are generated and combined to solve the problem. For Random Forest, each individual tree in a random forest spits out a class prediction and the class with the most votes is the model’s prediction

Multi-classs Classification

Multiclass classification model was used to try to predict for closed, operating and acquired companies. It showed that Industry Group is the most important feature. The important features are as follows:

```
cat_Industry_Group   Importance: 0.51

cat_Continent_Name   Importance: 0.08

cat_total_investment Importance: 0.06

cat_funding_rounds   Importance: 0.04

cat_diff_funding_year Importance: 0.04

cat_venture          Importance: 0.04
```

We used random grid search for hyper parameter tuning where

```
'n_estimators': 1600, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'sqrt',
'max_depth': 10, 'bootstrap': True
```

Using these parameter the accuracy rate came up to be 0.86

Target Status	Precision	Recall	F1-score	Support
Closed	0.00	0.00	0.00	456
Operating	0.86	1.00	0.93	7037
Acquired	0.00	0.00	0.00	663

Classification Report of Random Forest Multiclass Model

Binomial Classification

To predict for acquired and closed companies only, we use a random forest model too. According to the model the important features are:

```
cat_Industry_Group   Importance: 0.4

cat_total_investment Importance: 0.17

cat_Continent_Name   Importance: 0.08

cat_venture          Importance: 0.08

cat_funding_rounds   Importance: 0.04

cat_round_B          Importance: 0.04
```

We used random grid search for hyper parameter tuning where

```
'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'auto',
'max_depth': 10, 'bootstrap': True
```

Using these parameters the accuracy rate come up to be 0.69

Target Status	Precision	Recall	F1-score	Support
Closed	0.60	0.57	0.59	425
Acquired	0.74	0.76	0.75	680

Classification Report of Random Forest Multiclass Model

SVM

Support Vector Machine are a set of supervised learning methods used for classification, regression and outlier detection. All of these are common tasks in machine learning.

By using the SVM classification model, we achieved the best accuracy rate of 0.86 compared to all other multi-class classification models like Decision tree, Random Forest and KNN.

Below is the classification report for all target statuses.

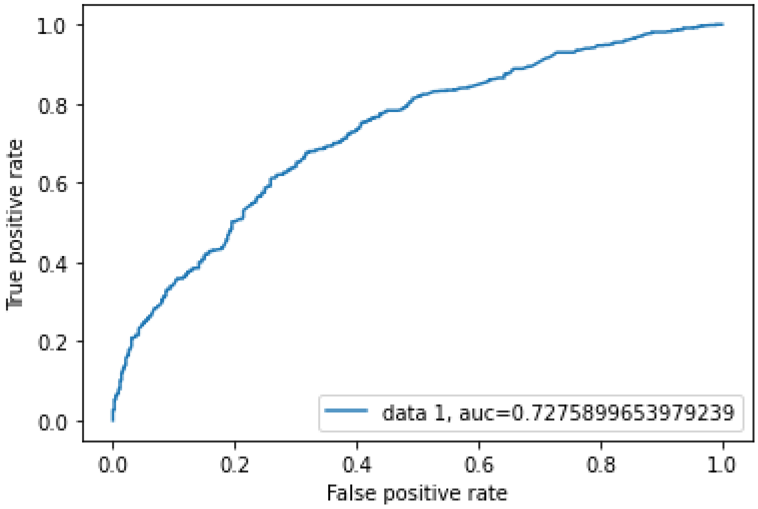
Target Status	Precision	Recall	F1-score	Support
Acquired	0	0	0	456
Operating	0.86	1	0.93	7037
Closed	0	0	0	663

Using grid search, we tried to find features that would get the best accuracy rate. Our grid search gave us almost the same results for fitting 5 folds for each of 9 candidates, totalling 45 fits. The accuracy rate came up to be 0.86

Logistic Regression

Logistic regression is a machine learning model that deals with binary target variables. It helps to explain the relationship between the binary target variable and the nominal, interval ,ordinal or ratio level independent variables. It's quite easy to explain the results of a logistic model to non-technical people.

ROC curve



Based on the ROC index Logistic Model was selected.It had 0.72 ROC index. A perfect ROC index is 1. Hence our ROC index of 0.72 means our model is performing well.

To achieve better results we tried fitting the model with both balanced and liblinear solvers. But we got the same accuracy as 0.73.

KNN

CONCLUSION

Industry, continent and total investment are important features. We received the best result when we used SVM and Random Forest for Multi Class Classification. Received the best result when we used Random Forest for Binomial Classification.

For future scope, we would like more data for closed and acquired companies, test model with one-hot encoding, test with other models like Naive Bayes and XG Boost, test with KNN and SVM on Binary Classification Model. Using Crunchbase API, we can also make a real time dashboard and deploy a model so that it can assist investors and founders.

REFERENCES

https://www.kaggle.com/arindam235/startup-investments-crunchbase

https://www.investopedia.com/articles/personal-finance/102015/series-b-c-funding-what-it-all-means-and-how-it-works.asp

https://www.investopedia.com/terms/s/startup.asp

https://support.crunchbase.com/hc/en-us/articles/115010458467-Glossary-of-Funding-Types

https://support.crunchbase.com/hc/en-us/articles/360043146954-What-Industries-are-included-in-Crunchbase-