

Machine-learning approach to start-up success



Table of contents

Seed: Introduction

A: Data sources

B: Data wrangling

C: Data story

D: Data models

Exit: Recommendations

Introduction

Forbes defines start-up as a company that is based on the “culture and mentality of innovating ... to solve critical pain points”*.

Given the fluid definition of a start-up and the difficulty of tracking all of them down, it is hard to tell exactly how many start-ups there are globally or even in the US at any given moment.

Many start-ups will fail before truly taking off, while many start-ups will fail after getting a significant amount of funding. On the other hand, many start-ups will succeed even with little funding.

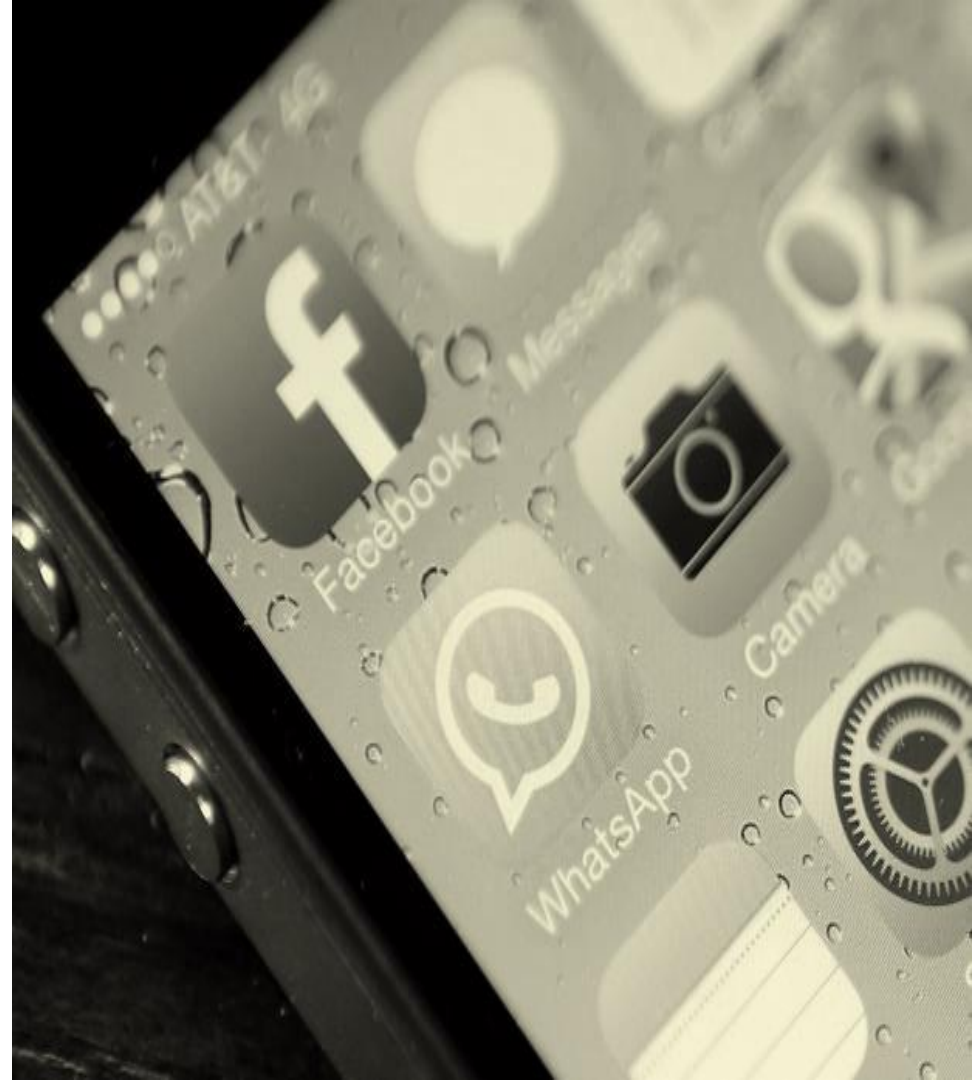
Even if not closing down, no company can exist in a start-up form forever. Driven by funding, start-ups will have to provide an exit for its shareholders. At that point, an acquisition or an IPO becomes inevitable.

* “What is a start-up” -- Forbes, December 16th 2013

Thesis:

If acquisition or IPO = success and
closing down = failure then:

Can we (as investors, for example) predict as
early as possible which start-up will succeed and
which one will fail?



Data source:

Crunchbase:

<https://www.crunchbase.com/>

Datasets used in the project were downloaded on December 4th 2015 and were made available on github

Use of this data is governed by the [CrunchBase Terms of Service and Licensing Policy](#)

The database is broken down into three main datasets:

For the purpose of this study we will be focusing on the first two.

Companies

Relevant information for each specific start-up

Data points:

- ☐ Company name/ company ID
- ☐ Status (operating, closed, IPO, acquired)
- ☐ Homepage url
- ☐ Sector
- ☐ Country/state/region/city
- ☐ Total funding obtained
- ☐ Total number of funding rounds
- ☐ Date founded
- ☐ Date of first funding round
- ☐ Date of last funding round

Investments

Breakdown of funding rounds for each start-up

Data points:

- ☐ Company name/company ID
- ☐ Sector
- ☐ Country/state/region/city
- ☐ Investor name
- ☐ Investor sector
- ☐ Investor country/state/region/city
- ☐ Funding round type (e.g. Series A)
- ☐ Date of funding round
- ☐ Amount raised in the round

** This dataset has multiple lines for each start-up to account for each investor in each funding round*

Acquisitions

Information on acquisitions by start-ups and of start-ups

Data points:

- ☐ Company name/company ID
- ☐ Sector
- ☐ Country/state/region/city
- ☐ Bidder name/bidder ID
- ☐ Bidder sector
- ☐ Bidder country/state/region/city
- ☐ Date acquired
- ☐ Price/currency

Data wrangling

Company
duplicates

Missing values

Converting
date variables
from string to
date format

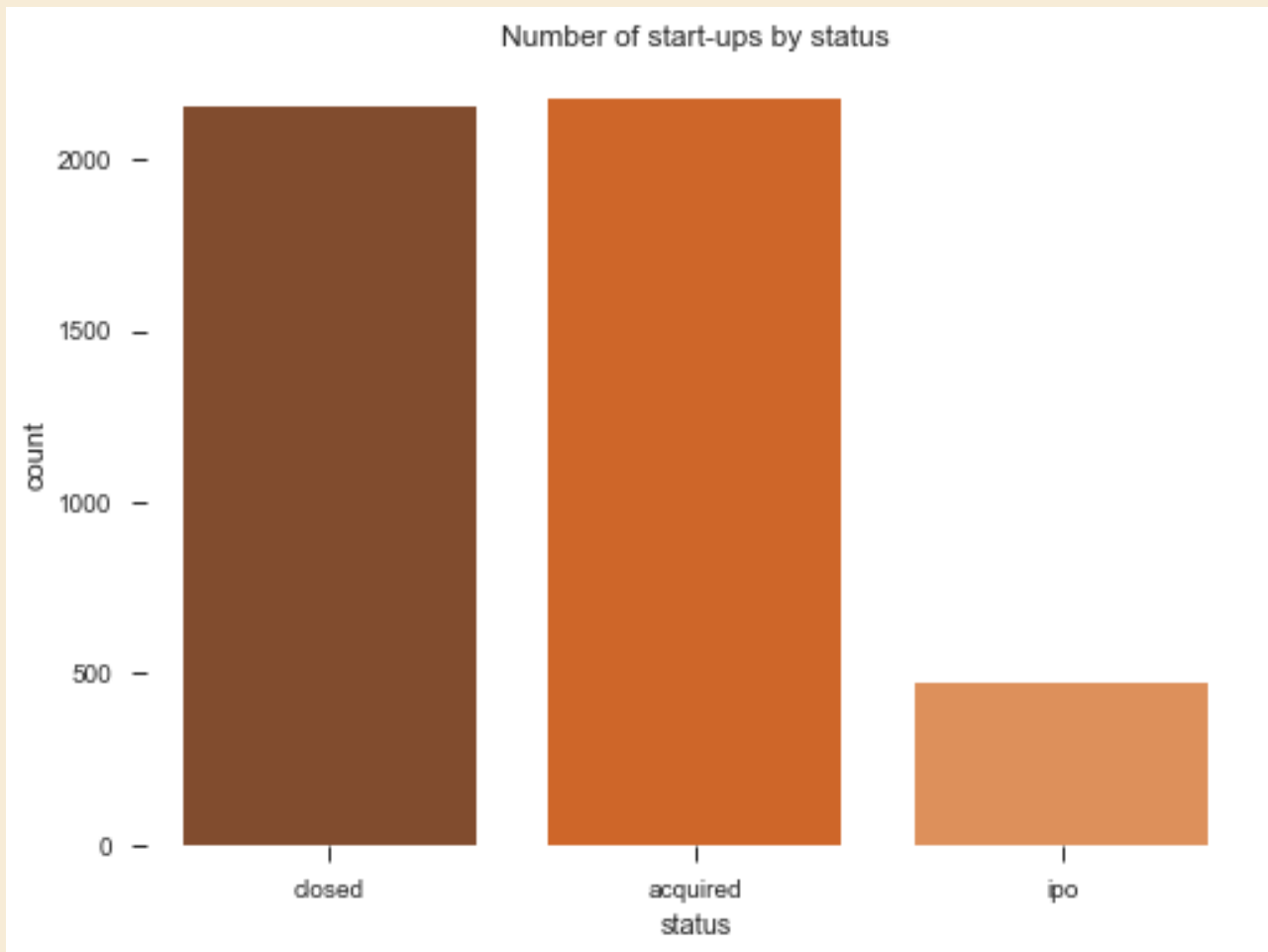
Merging
datasets

Data story

It is often said that only 1% of all ideas succeed.

However, Crunchbase data paints a different story. It shows that more than 50% of companies with at least one funding round succeed.

*Assuming that there is no selection bias within Crunchbase data

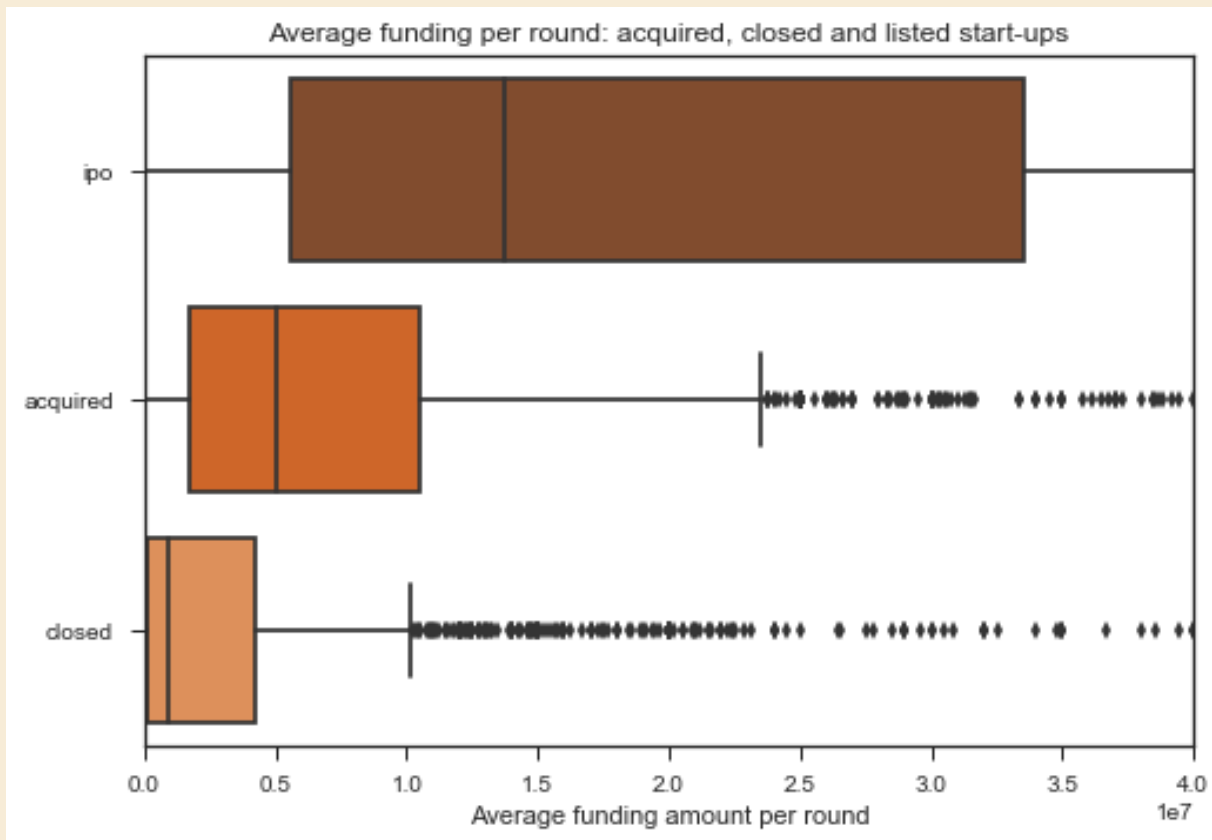


Data story

The data also reveals some differences between successful and unsuccessful companies.

On average, the latter were involved in fewer funding rounds and received less funding per round.*

*This is to be expected: funding amount generally grows in each subsequent funding round.

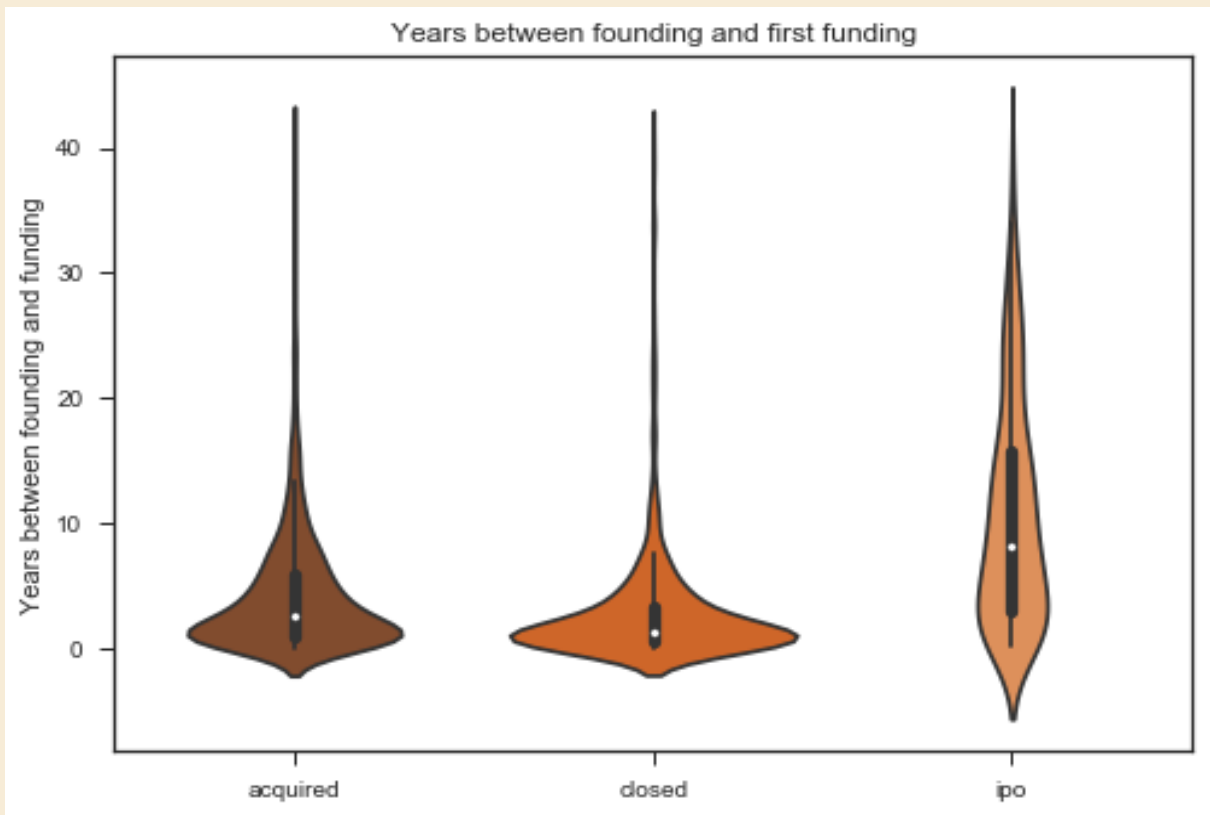


	Listed	Acquired	Closed
Average num. of funding rounds	2.6	1.9	1.4

Data story

Finally, unsuccessful start-ups seem to rush to get their first funding, while successful start-ups wait longer.

But even if there does appear to be a difference between start-ups early on, can we actually predict whether a start-up will fail or succeed?



	Listed	Acquired	Closed
Median number of years until first funding	8.4	2.5	1.3

Testing data models

Prior to testing different models to predict start-up success, the following features were selected.

Each row was represented by a start-up.

The response variable was binary:

0 for unsuccessful/closed companies;

1 for successful/listed or acquired companies



Investor information

Transformed to make it readable by the model (explained in the next slide)



Start-up sector, country, state, region, city

Converted to dummy variables



Total funding raised

Normalized



Number of funding rounds



Year founded

Normalized



Days between founding and first funding

Normalized

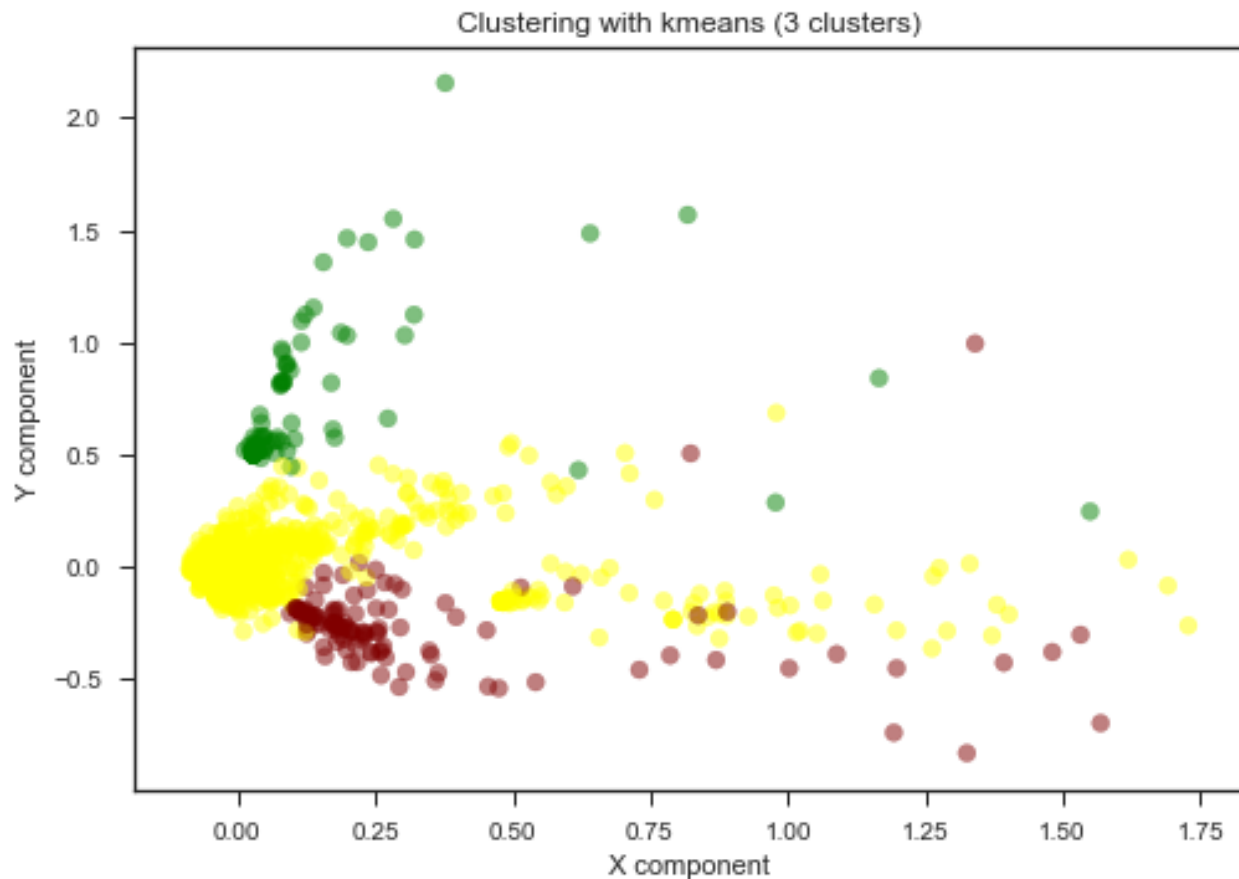
Plus, average funding per round

Testing data models

Investor information was transformed by turning each investor into a separate column. The values in the resulting dataset represented the number of funding rounds a specific investor participated in for a specific company.

Because investor data amounted to more than 4500 columns, to avoid the curse of dimensionality (and to speed up the model), PCA was used to reduce dimensionality to 600 principal components.

The first two components are clustered through kmeans here:



Selecting the final data model

Logistic Regression and Random Forest were cross-validated on 40% of the data and validated using 30% of the data. The combined 70% were used for final training, with the remaining 30% used for final testing.

Based on results provided here, LR was selected as the final prediction model, with accuracy in the testing phase = **0.694**

Cross-validation -> Validation -> Training -> Testing

Logistic Regression		Random Forest	
Accuracy:	0.675	Accuracy:	0.638
Sensitivity:	0.852	Sensitivity:	0.934
Specificity:	0.415	Specificity:	0.205

Baseline accuracy: 0.594

Recommendations

1. While this model can be used as early as Series A funding round to identify potential successful and unsuccessful candidates, **the more funding rounds a company has gone through, the more information there is for the model.**
2. **Investors are encouraged to consider individual features**, such as days/years before the company got its first funding, selection of investors in each round, or average amount of funding per round, **to raise concerns about a potential investment.**
3. It **would be ideal to invest into acquisition of further start-up data**, e.g. information on founders, start-up financial information, or start-up PR and marketing information, **to improve performance of the model.**

Further study

Features:

- Use the most updated Crunchbase data, ideally real-time data, for the model
- Utilize the acquisitions dataset to generate further features
- Invest into acquisition of additional, early-stage start-up data

Model:

- Test other models (aside from RF and LR) or a combination of different models
- Predict not only success/failure but additional company statuses (IPO/acquisition). The latter might be of interest not just to investors, but to the advisory community as well.



Thank you!