
# PREDICTION OF SUCCESS IN EARLY-STAGE STARTUPS USING MACHINE LEARNING

Javier Arroyo
Juan A. Recio
Guillermo Jiménez

Departamento de Ingeniería del Software e Inteligencia Artificial
Universidad Complutense de Madrid

# Venture Capital and Machine Learning

- Venture Capital (VC) attracts investors looking for financial returns and innovation sprints
- A good opportunity still for investments in early-stage projects
- Data-driven investors are typical in hedge-funds but not in VC
- Machine learning can help to identify patterns or signals of potential success
- Platforms like Crunchbase or Pitchbook gather data from thousands of companies

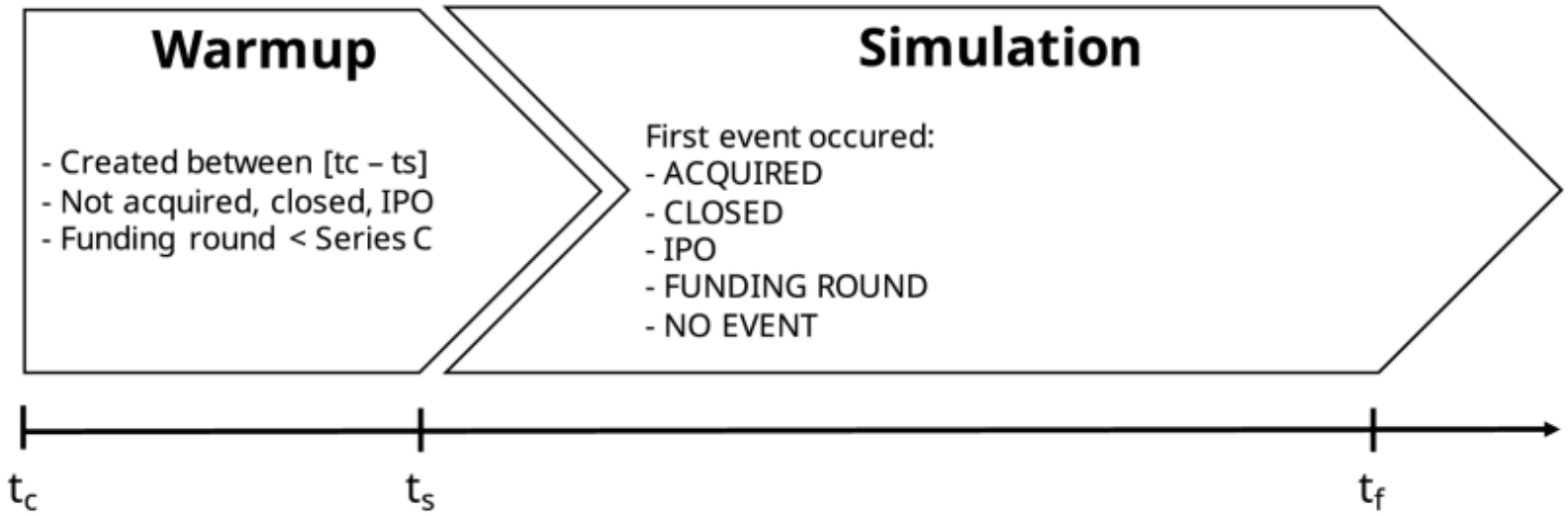
# State of art of data-driven VC

- Target variable is binary (Success or not)
  - IPO or Acquisition
- Logistic Regression is typically used
  - ML methods are not so popular yet
- Small samples of companies
  - Typically hundreds or few thousands
  - If bigger samples are used, the time-windows is not realistic

# Our approach

- Focused on early-stage companies
  - Earlier than series C
- Time-aware analysis
  - Considering a realistic time window (3 years)
- Multi-class target variable
  - First event achieved by the company in the time window
    - IPO, Acquired (AC), New funding round (FR)
    - Closed (CL), No Event (NE)

# Our approach




Aug 2011 Aug 2015 Aug 2018

Startups have a high failure rate in the early years;
e.g., at four years was about 44 percent in the US

# Data sample

- 120,507 companies from Crunchbase
- Predictor variables (105)
  - Company information:
    - Profile: country, age in months, sector (46 categories)
    - Data available in Crunchbase (email, linkedIn, facebook, twitter)
  - Funding information
    - Data about the number of funding rounds before $t_s$
    - Data about the last funding round
    - Data about investors
  - Founders information (*bad quality*)
    - Number of founders (males and females), number of countries of origin, number of degrees, …

# Time-aware simulation

- We removed all *dated* data with a date before the start of the simulation window $t_s$
- We did not consider data that could have changed during the simulation window (number of employees, managers of the company…)

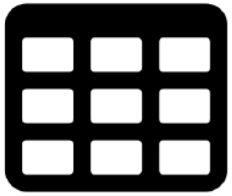
# Our approach

Crunchbase companies

Earlier than Series C
< 4-years old in August 2015

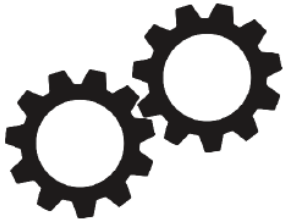
*3-year simulation window*

Predictor Variables



Company information
Funding information
Founders information

Machine Learning



Support Vector Machines
Gradient Tree Boosting
Extremely Randomized Trees
Random Forests
Decision Trees

Target Variable

Acquired
IPO
Funding Round
No Event
Closed

# Data sample

- Target variable distribution

| Class | Frequency | Ratio |
|---|---|---|
| CLOSED (CL) | 686 | 0.57% |
| ACQUIRED (AC) | 3,293 | 2.73% |
| FUNDING ROUND (FR) | 21,682 | 17.99% |
| IPO (IP) | 143 | 0.12% |
| NO EVENT (NE) | 94,703 | 78.58% |

"No event" probably means "Closed"
or an uninteresting company in any case

# Machine learning methods

- Methods considered
  - Decision Tree
  - Random Forest
  - Extremely Randomized Tree
  - Gradient Tree Boosting
  - SVM

} Feature selection
Feature importance

- K-fold cross validation (k=5)

# Global results

| Classifier | Accuracy |
|---|---|
| Decision Trees | 74.6 |
| Random Forests | 81.8 |
| Extremely Randomized Trees | 81.9 |
| Gradient Tree Boosting | **82.2** |
| Support Vector Machines | 81.7 |

# Results per class

| Class | Frequency | Ratio |
|---|---|---|
| CLOSED (CL) | 686 | 0.57% |
| ACQUIRED (AC) | 3,293 | 2.73% |
| FUNDING ROUND (FR) | 21,682 | 17.99% |
| IPO (IP) | 143 | 0.12% |
| NO EVENT (NE) | 94,703 | 78.58% |

| | CL | | | NE | | |
|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| DT | 0.02 | 0.02 | 0.02 | 0.85 | 0.85 | 0.85 |
| RF | 0.00 | 0.00 | 0.00 | 0.85 | 0.94 | 0.89 |
| ERT | 0.00 | 0.00 | 0.00 | 0.85 | 0.94 | 0.89 |
| GTB | 0.00 | 0.00 | 0.00 | 0.85 | 0.95 | 0.90 |
| SVM | - | - | - | 0.83 | 0.96 | 0.89 |

| | FR | | | AC | | | IP | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| DT | 0.43 | 0.43 | 0.43 | 0.09 | 0.10 | 0.10 | 0.04 | 0.03 | 0.04 |
| RF | 0.60 | 0.44 | 0.51 | 0.33 | 0.03 | 0.05 | 0.44 | 0.03 | 0.05 |
| ERT | 0.61 | 0.43 | 0.51 | 0.31 | 0.03 | 0.05 | 0.27 | 0.02 | 0.04 |
| GTB | 0.64 | 0.40 | 0.49 | 0.17 | 0.003 | 0.01 | 0.07 | 0.01 | 0.02 |
| SVM | 0.64 | 0.33 | 0.44 | 0.00 | 0.00 | 0.00 | - | - | - |

Recall is not critical, if high enough

# Error reintepretation

- Classification errors may not be real (or harmful) errors for a VC investor
  - If FR is predicted, but the company is Acquired
- We create two groups:
  - Good results: IPO, Acquired, Funding Round
  - Bad results: Closed and No Event

# Error reintepretation

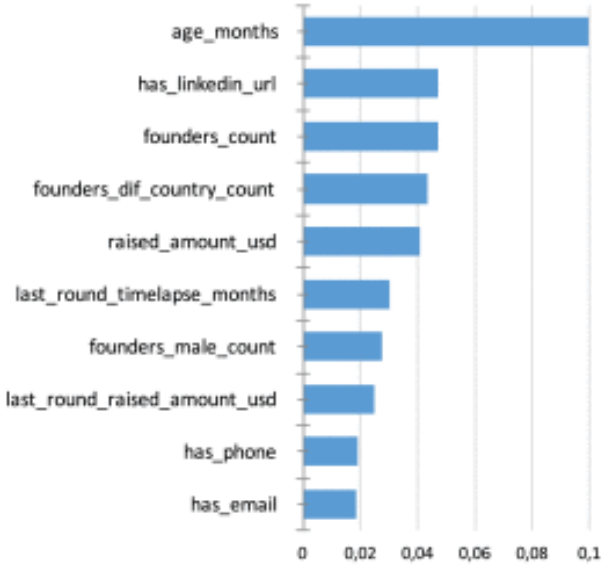
| Observed target \ Predicted Target | NO_EVENT | ACQUIRED | CLOSED | IPO | FUNDING_ROUND |
|---|---|---|---|---|---|
| NO_EVENT | 88830 | 137 | 1 | 4 | 5731 |
| ACQUIRED | 2537 | 97 | 0 | 0 | 659 |
| CLOSED | 567 | 2 | 0 | 0 | 117 |
| IPO | 102 | 1 | 0 | 4 | 36 |
| FUNDING_ROUND | 11999 | 57 | 2 | 1 | 9623 |

# Reinterpreted results

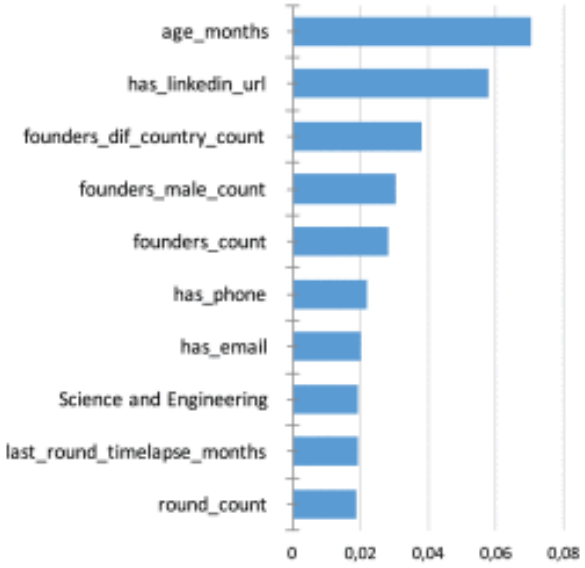
| Classifier | Bad | Good |
|---|---|---|
| Decision Trees | 0.86 | 0.45 |
| Random Forests | 0.86 | 0.64 |
| Extremely Randomized Trees | 0.86 | 0.64 |
| Gradient Tree Boosting | 0.85 | 0.68 |
| Support Vector Machines | 0.84 | 0.68 |

| | Bad | | | | Good | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CL | | NE | | FR | | AC | | IP | |
| | Prec. | TP | Prec. | TP | Prec. | TP | Prec. | TP | Prec. | TP |
| Decision Trees | 0.74 | 535 | 0.86 | 80638 | 0.47 | 10195 | 0.34 | 1228 | 0.39 | 53 |
| Random Forest | 0.33 | 1 | 0.86 | 89397 | 0.64 | 10318 | 0.53 | 155 | 0.56 | 5 |
| Extremely Randomized Trees | 0.33 | 1 | 0.86 | 89721 | 0.65 | 10039 | 0.47 | 128 | 0.55 | 6 |
| Gradient Tree Boosting | 0.44 | 24 | 0.85 | 91016 | 0.68 | 9139 | 0.55 | 35 | 0.41 | 11 |
| Support Vector Machines | - | 0 | 0.84 | 91817 | 0.68 | 7604 | 1 | 1 | - | 0 |

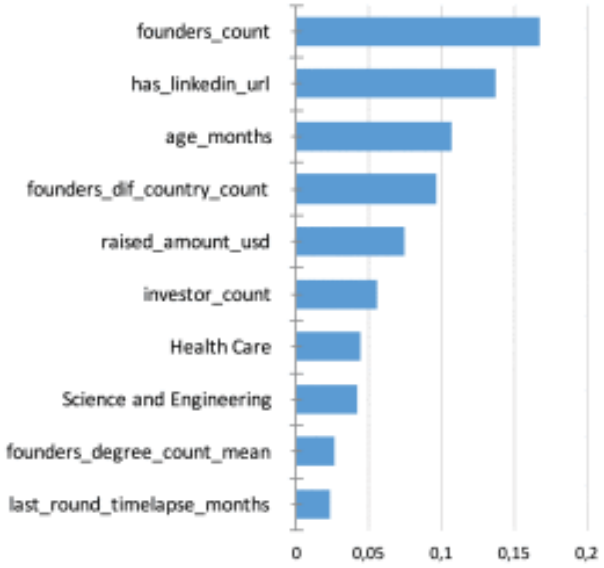
# Feature importance




Random forests




Extremely randomized trees




Gradient tree boosting

- Some variables represent how complete is the company profile in Crunchbase
- Others identify attractive business sectors or countrie

# Potential biases

- Successful companies tend to have more data in their Crunchbase profiles
  - As they progress, they complete their Crunchbase profile (e.g. founders data)
- Our data was retrieved at the end of the simulation window
  - We would need an image of the database from the start of the simulation
  - In our data, we do not know which data was added during the simulation

# Conclusions

- The succesful rates of our approach are extremely good (7 out of 10)
  - VCs often expect 1 or 2 out of 10 succesful rate (with outsized results)
- Our multi-class target variable makes possible to focus on different levels of “risks” and “rewards“
  - Different portfolio strategies are posible
- Higher data quality should produce better results
  - e.g. about founders and managers

# Thank you!

Thank you!

## For those interested:

J. Arroyo, F. Corea, G. Jiménez-Díaz, J. A. Recio-García (2019)
*Assessment of machine learning performance for decision support in venture capital investments*
**IEEE Access** 7, 124233-124243