



Predicting Success Rate of Startups

Presentation by: Prianka Ball, Kalyani Pavuluri & Shivprasad Chavan



Problem Statement

- Startups have very high failure rate.
- People who invest in startup like angel investors or venture capitalists have a very high risk while investing in startups.





Goal

- Find important features that lead to startup success
- Forecasting a company's success with supervised machine learning methods

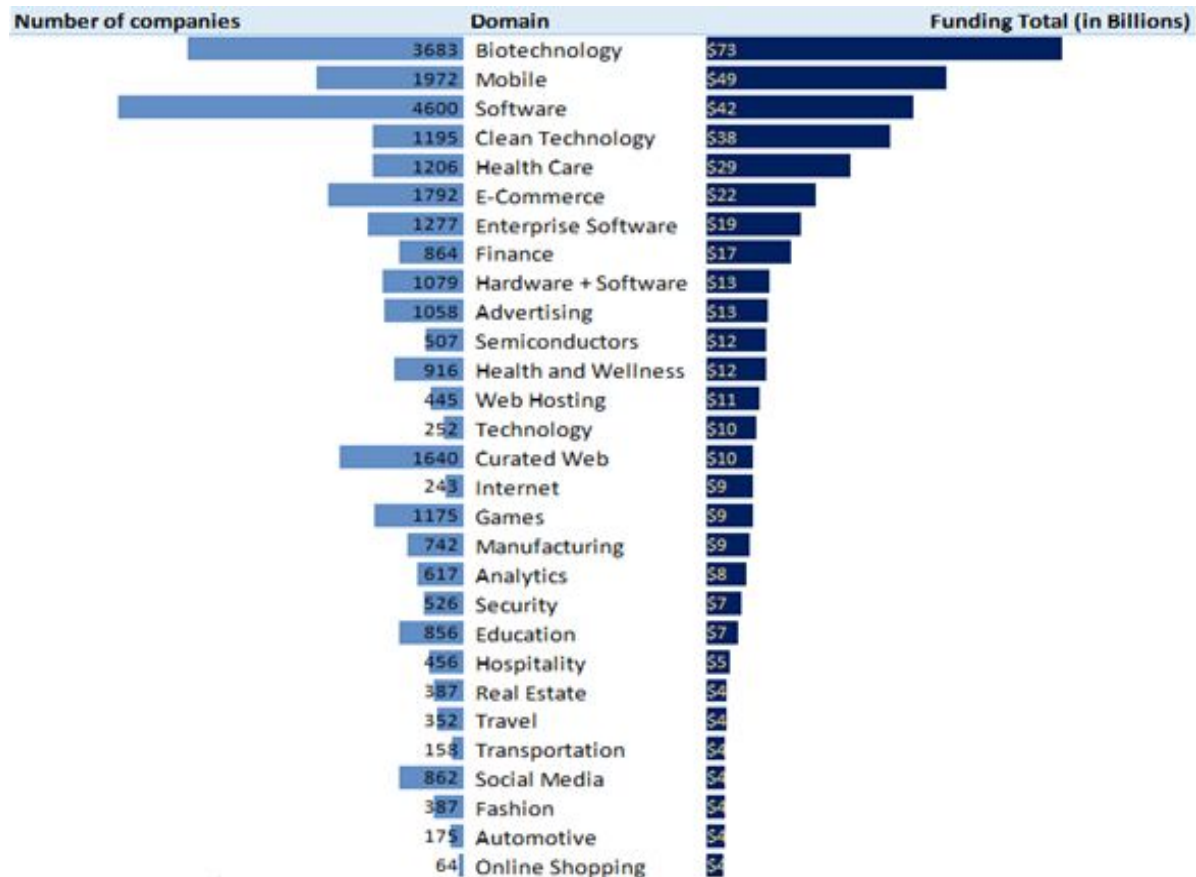


Approach

- Exploratory Data Analysis
- Feature Engineering
- Testing Models
- Hyperparameter Tuning

Exploratory Data Analysis

Relationship between number of companies under each domain and total amount of funding received by that particular domain.





Exploratory Data Analysis

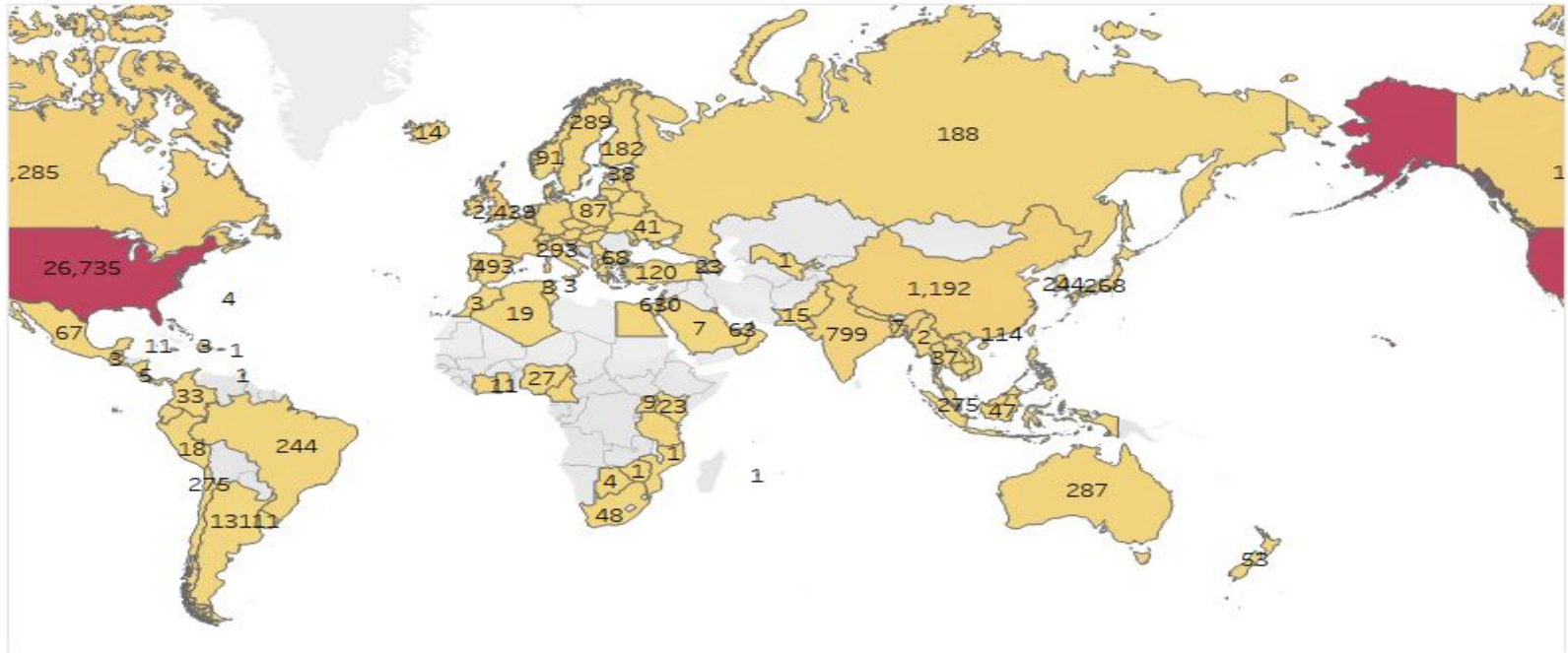
23,278 companies were able to achieve venture capital and 13,841 companies were able to achieve seed capital.

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

Exploratory Data Analysis

Success companies in different locations





Cleaning Data/Pre-Processing steps

- Remove space from columns
- Remove characters like , or - from the data
- Turning data into date
- Turning data into number



Feature Engineering

- Difference between last funding and first funding day in days, years and months
- Calculating total investment.
- Remove null values.
- Group 736 market values into 43 industry groups.
- Group 113 countries to 5 continents.
- Turning all numerical values into categories based on their spread.
- Label Encoding
- Turned columns that has a lot of zeros into categorical values based on if they received that specific type of funding or not.



Techniques and Algorithms

- Upsampling and Downsampling
- Multi Class Classification
- Binomial Classification Model
- Models Used:
 - Decision Tree
 - Random Forest
 - KNN
 - SVM
 - Logistic Regression (Only for Binomial Classification)
- Grid Search and Random Search for Hyper Tuning model parameters



Model Output: Multi-Class Classification

Got the best result when we used downsampling

Model	Accuracy Rate
Decision Tree	0.57
Random Forest	0.59
KNN	0.55
SVM	0.61

Classification Report from SVM Model

	precision	recall	f1-score	support
0	0.08	0.05	0.06	456
1	0.89	0.64	0.75	7037
2	0.14	0.61	0.23	663
accuracy			0.61	8156
macro avg	0.37	0.43	0.35	8156
weighted avg	0.79	0.61	0.67	8156



Model Output: Binomial Classification

Classification Report from Random Forest Model

Model	Accuracy Rate
Decision Tree	0.69
Random Forest	0.70
Logistic Regression	0.69

	precision	recall	f1-score	support
Class 1	0.61	0.57	0.59	425
Class 2	0.74	0.77	0.76	680
accuracy			0.70	1105
macro avg	0.68	0.67	0.67	1105
weighted avg	0.69	0.70	0.69	1105



Model Conclusion

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



Future Scope

- More data for closed and acquired companies.
- Access to Crunchbase API for real time data.
- Testing model with one-hot encoding.
- Testing with other models like Naive Bayes and XG Boost. Test with KNN and SVM on Binary Classification Model.
- Making real time dashboard and deploying model so that it can assist investors and founders