Model Selection:
Team 1: Simple/Deep Neural Network, Naive Rule
Team 2: Classification tree, XGBoost
Team 3: kNN,
Team 4: Logistic Regression, Random Forest

Data Cleaning

Leave log on any column processing
(Handling/Removing NA, Create Dummy Variable, Split Text, Categorize Variable)
Team 1:
(For Example: “Industries”: Split Text, create dummy variable from the split)

Team 2:

Team 3:

Team 4:

Proposal for Data Cleaning

4/9/2020 Yida

IPO. Date: maybe we can remove it too? Since a large proportion of those companies are private so there are too many missing values that may affect the accuracy. Plus, the IPO dates do not make sense to private companies.

Acquisition. Status: I think this variable is interesting and should have some impact on the CB rank, but there are not a lot of companies that have this kind of activity. How do you guys want to deal with this variable?

Number. of. Events: If the number of events is not available to the public, can we assume that there is no impact at all? And we can just put those NA values to 0 and make the variable as a numeric one.

Number. of. Lead. Investors: I am confused about how to deal with it. Too many missing values and the range is so small.

IPqwery...Total.Patents & IPqwery...Total.Trademarks: Transfer NA to 0?

Dependent Variable (1)

CB rank (Categorical or Numeric)

Independent Variable before Cleaning (16)

	Column Name		Team Assignee
1	Industries (Categorical)	Create dummy variables for each industry, such as finance, commerce, technology, etc. Identify each company's main business and categorize them into major industries. E.g. Hulu, a film & TV company, can be classified into entertainment industry.	(deleted)
2	Headquarters. Location (Categorical)		
3	Estimated. Revenue. Range (Categorical)	Since there are a lot of missing values, we can transfer them to ranges and create dummy variables.	Team 1
4	Founded. Date (Numeric)	Transfer those dates to days founded.	Team 4 Finished - Transformed to Days_Founded (as of 4/12/2020)
5	Industry. Groups (Categorical)	We can keep either this variable or Industries? The two variables are almost identical.	Team 1 Finished Using the first group

6	Number. Of. Founders (Numeric)	There are some missing values but not too many, so we may consider it as a numeric variable with continuous values.	
7	Number. of. Employees	Set different ranges as it already did and create dummy variables.	Team 1
8	Number. Of. Funding. Rounds (Numeric)	Keep it as original values.	
9	Total. Funding. Amount (Numeric)	Transfer the data in 1,000s.	Team 1
10	Number. Of. Investors (Numeric)	Similar to *Number. Of. Founders.*	
11	IPO. Status (Categorical)	Create dummy variables.	Team 4 Finished
12	Funding. Status (Categorical)	Create dummy variables.	Team 4 Finished
13	Last. Funding. Date (Numeric)	Refer to *Founded. Date.*	Team 4 Finished - Transformed to Days_after_Last_Funding
14	Last. Funding.	Refer to *Total. Funding. Amount.*	Team 4

	Amount. Currency.. in.USD (Numeric)		Finished
15	Last. Funding. Type (Categorical)	Create dummy variables.	Team 4 Finished

Output/ Not variables		
Y		
id		
Organization.Name		
CB.Rank..Company.		
Description		
Categorical Variables		
Headquarters.Location		
Industry.Groups	Team 1	✓
Estimated.Revenue.Range	Team 1	
Founded.Date		
Number.of.Employees	Team 1	
Number.of.Funding.Rounds		
Total.Funding.Amount	Team 1	
IPO.Status	Team 4	✓
IPO.Date		
Acquisition.Status		
Funding.Status	Team 4	✓
Last.Funding.Type	Team 4	✓
Last.Funding.Date		
Numerical Variables		
Number.of.Founders		
Number.of.Investors		
Number.of.Events		
Number.of.Lead.Investors		
Last.Funding.Amount.Currency..in.USD.	Team 4	✓
IPqwery...Total.Patents		
IPqwery...Total.Trademarks		