# What Makes a Startup Successful?

Taylor Niedzielski, Noah Suttora, Sonia Patel, Julia Nelson

# Start-Up Companies

The Problem:

Startups can be costly from investment money and time commitment. There is confusion about which attributes can help or hurt a new company.

# Start-Up Success Rate

Objective:

Discover factors that positively impact and negatively impact success of American start-up companies.

# Process to Meet Objective:

1) Get **data**
2) Extract **information** via exploratory data analysis
3) Gain **knowledge** using machine learning algorithms
4) Acquire **wisdom** to understand optimal actions

# 1.) DATA

# The Dataset

The data collected is from the years **1999-2013** with only American based companies.

Originally had 116 different attributes that we narrowed down to **43** attributes.

We narrowed down this list by:

- Usability of attribute
- Understandability of the attribute
- Relevance of attribute
- > 35% of attribute data missing

# Cleaning The Dataset

- The attributes with missing values were propagated with the mean, if they were a numerical type, or mode, if they were a categorical type.

- We shortened the attribute names for clarity and conciseness.

# Qualitative (Categorical) Data Variables

- TeamSizeGrowth
- TopCompanyExp
- StartupExp
- SuccessfulStartupExp
- Big5Partner
- ConsultingExp
- HighestEducation
- Fortune100Exp
- Fortune 500Exp
- Fortune1000Exp
- FortuneExp
- Focus Functions
- ProductorService
- DataFocus
- ConsumerDataFocus
- DataStructureFocus
- SubscriptionBased
- CloudPlatformBased
- LocalGlobal
- BusinessModel
- CapitalIntensive
- CrowdsourcingBased
- CrowdfundingBased
- B2BorB2C
- GlobalExposure
- PricingStrategy
- HyperLocalisation,
- LongtermFounderRelationship
- RecessionalSurvival

# Quantitative (Numeric) Data Variables

- FoundingYear
- Age
- NumSeedInvestors
- NumAngelorVCInvestors
- NumFounders
- NumAdvisors
- GooglePageRank
- NumDirectCompetitors
- LastFundingRoundAmount
- SeniorLeadershipTeamSize
- EmployeesPerYear
- NumFounderRecognition

# 2.) EXTRACT INFORMATION

## Exploratory Data Analysis via Linear Regression

# Correlation Coefficient (r) and Strength

| \|r\| | Strength |
|---|---|
| 0.70 - 1.00 | Very Strong |
| 0.50 - 0.69 | Strong |
| 0.30 - 0.49 | Moderate |
| 0.10 - 0.29 | Weak |
| 0.01 - 0.09 | Very Weak |

# No Correlation and Very Weak Correlation Variables

**No Correlation (r = 0.0 or NA)**

- DataStructureFocus (+0.008)
- NumFounderRecognition (NA)

**Very Weak Correlation (r = 0.01 - 0.09)**

- CapitalIntensive (-0.02)
- SuccessfulStartupExp(+0.02)
- GlobalExposure (+0.02)
- NumSeedInvestors(+0.03)
- BusinessModel (+0.03)
- NumFounders (+0.03)
- NumAngelorVCInvestors(+0.05)
- LastFundingRoundAmount(-0.04)

- PricingStrategy (-0.06)
- StartupExp (+0.06)
- HyperLocalisation (-0.07)
- ProductorService (-0.09)
- Big5Partner (+0.09)
- DataFocus (+0.09)
- TopCompanyExp (+0.09)

# Weak Correlation Variables (r = 0.10 - 0.29)

- SubscriptionBased (+0.10)
- TopCompanyExp (+0.10)
- CloudPlatformBased (-0.10)
- NumDirectCompetitors (-0.10)
- CrowdsourcingBased (-0.12)
- CrowdfundingBased (0.12)
- HighestEducation (0.15)
- Fortune100Exp (+0.16)
- Fortune500Exp (+0.16)
- Age (-0.17)
- EmployeesPerYear (+0.17)
- ConsumerDataFocus (+0.17)
- ConsultingExp (-0.19)
- Founding year (+0.19)
- NumAdvisors (+0.19)
- SeniorLeadershipTeamSize (+0.19)
- FortuneExp (+0.20)
- Fortune1000Exp (+0.21)
- LongtermFounderRelationship (+0.23)
- GooglePageRank (-0.26)

# Moderate and Very Strong Correlation Variables

**Moderate Correlation (r = 0.30 - 0.49)**

- B2BorB2C (-0.30)
- LocalGlobal (-0.33)

**Very Strong Correlation (r = 0.70 - 1.00)**

- RecessionSurvival (+0.73)

# 3.) GAIN KNOWLEDGE

## Machine Learning Algorithms

# Algorithms We Used

k-Nearest Neighbors (kNN)

Naive Bayes (NB)

Classification and Regression Tree (CART)

Random Forest (RF)

C5.0 Classification (C50)

Artificial Neural Network (ANN)

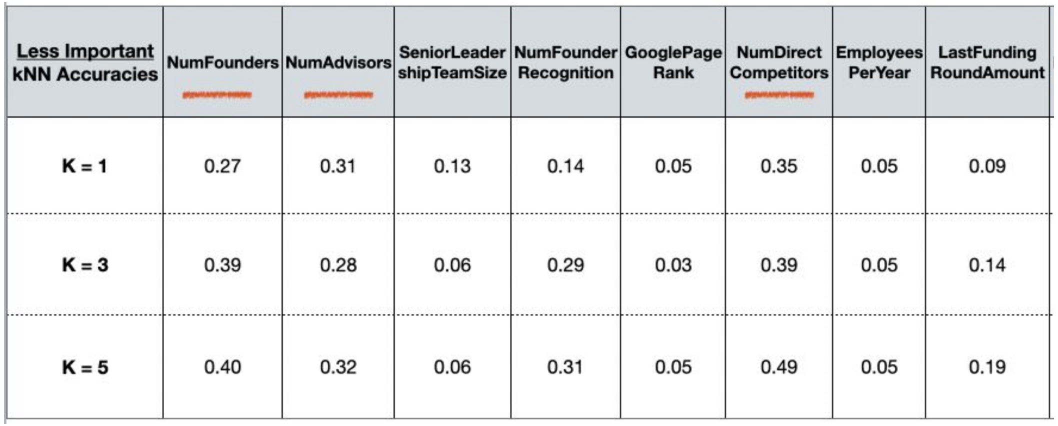
# k-Nearest Neighbors

# k-Nearest Neighbors

Applied on 11 variables

3 variables seem to play an important role in a company's success

| Important kNN Accuracies | Age | NumSeedInvestors | NumAngelorVCInvestors |
|---|---|---|---|
| K = 1 | 0.73 | 0.71 | 0.62 |
| K = 3 | 0.59 | 0.65 | 0.68 |
| K = 5 | 0.57 | 0.63 | 0.70 |



| Less Important kNN Accuracies | NumFounders | NumAdvisors | SeniorLeadershipTeamSize | NumFounderRecognition | GooglePage Rank | NumDirect Competitors | Employees PerYear | LastFunding RoundAmount |
|---|---|---|---|---|---|---|---|---|
| K = 1 | 0.27 | 0.31 | 0.13 | 0.14 | 0.05 | 0.35 | 0.05 | 0.09 |
| K = 3 | 0.39 | 0.28 | 0.06 | 0.29 | 0.03 | 0.39 | 0.05 | 0.14 |
| K = 5 | 0.40 | 0.32 | 0.06 | 0.31 | 0.05 | 0.49 | 0.05 | 0.19 |

Average accuracy is about 65-70%...

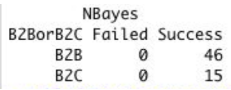
We expect future models will be minimum 70-85% accurate

# Naïve Bayes

# Naive Bayes

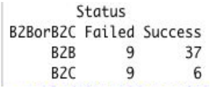
## B2BorB2C

**Predicted**



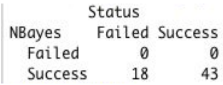
```
         NBayes
B2BorB2C Failed Success
     B2B      0      46
     B2C      0      15
```

**Actual**



```
        Status
B2BorB2C Failed Success
     B2B      9      37
     B2C      9       6
```

**Compared**

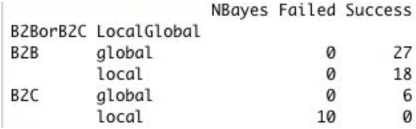


```
         Status
NBayes    Failed Success
  Failed       0       0
  Success     18      43
```

*NB Error rate: 0.30*

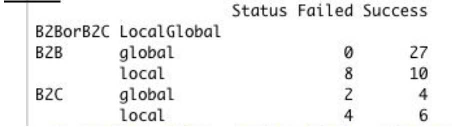
## B2BorB2C + LocalGlobal

**Predicted**



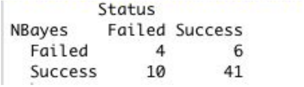
```
                     NBayes Failed Success
B2BorB2C LocalGlobal
B2B      global                  0      27
         local                   0      18
B2C      global                  0       6
         local                  10       0
```

**Actual**



```
                    Status Failed Success
B2BorB2C LocalGlobal
B2B      global                 0      27
         local                  8      10
B2C      global                 2       4
         local                  4       6
```

**Compared**

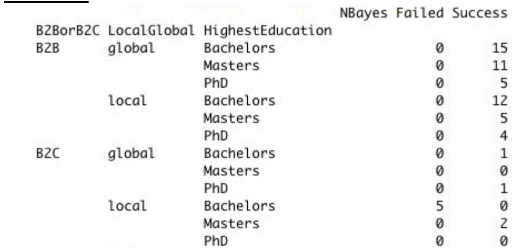


```
         Status
NBayes    Failed Success
  Failed       4       6
  Success     10      41
```

*NB Error rate: 0.26*

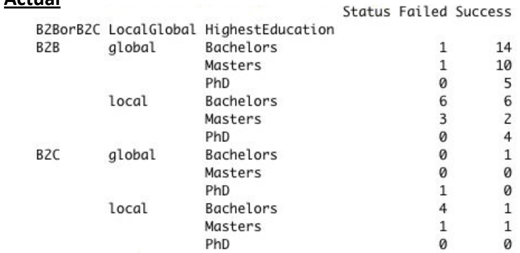
## B2BorB2C + LocalGlobal + HighestEducation

**Predicted**



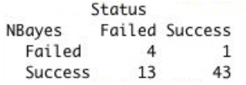
```
                                     NBayes Failed Success
B2BorB2C LocalGlobal HighestEducation
B2B      global      Bachelors                   0      15
                     Masters                     0      11
                     PhD                         0       5
         local       Bachelors                   0      12
                     Masters                     0       5
                     PhD                         0       4
B2C      global      Bachelors                   0       1
                     Masters                     0       0
                     PhD                         0       1
         local       Bachelors                   5       0
                     Masters                     0       2
                     PhD                         0       0
```

**Actual**



```
                                     Status Failed Success
B2BorB2C LocalGlobal HighestEducation
B2B      global      Bachelors                   1      14
                     Masters                     1      10
                     PhD                         0       5
         local       Bachelors                   6       6
                     Masters                     3       2
                     PhD                         0       4
B2C      global      Bachelors                   0       1
                     Masters                     0       0
                     PhD                         1       0
         local       Bachelors                   4       1
                     Masters                     1       1
                     PhD                         0       0
```

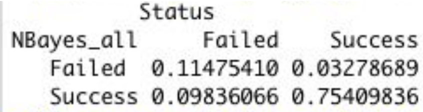
**Compared**



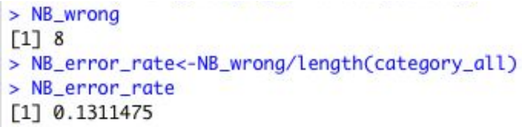
```
         Status
NBayes    Failed Success
  Failed       4       1
  Success     13      43
```

*NB Error rate: 0.23*

## All Categories



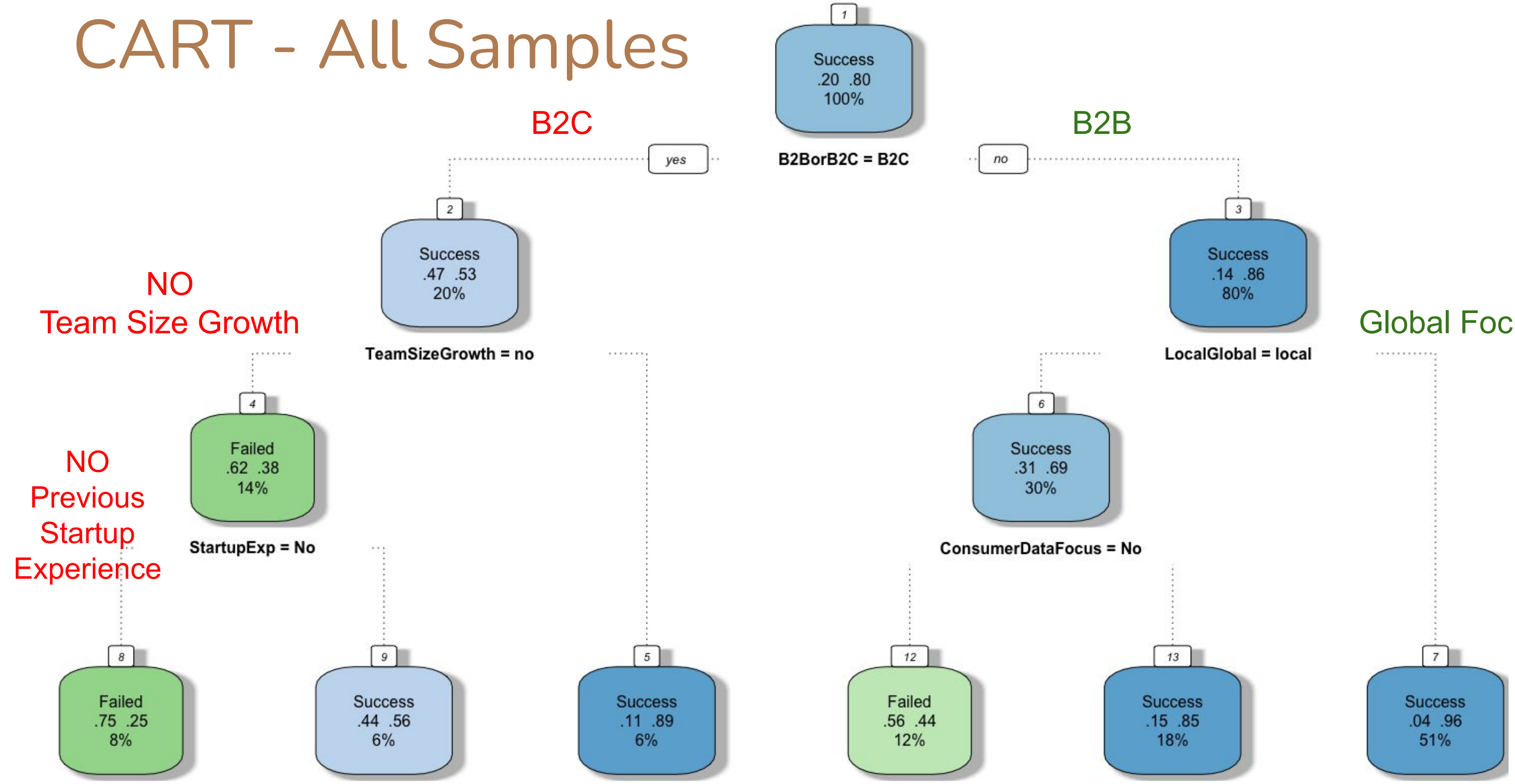
```
            Status
NBayes_all      Failed    Success
   Failed  0.11475410 0.03278689
   Success 0.09836066 0.75409836
```



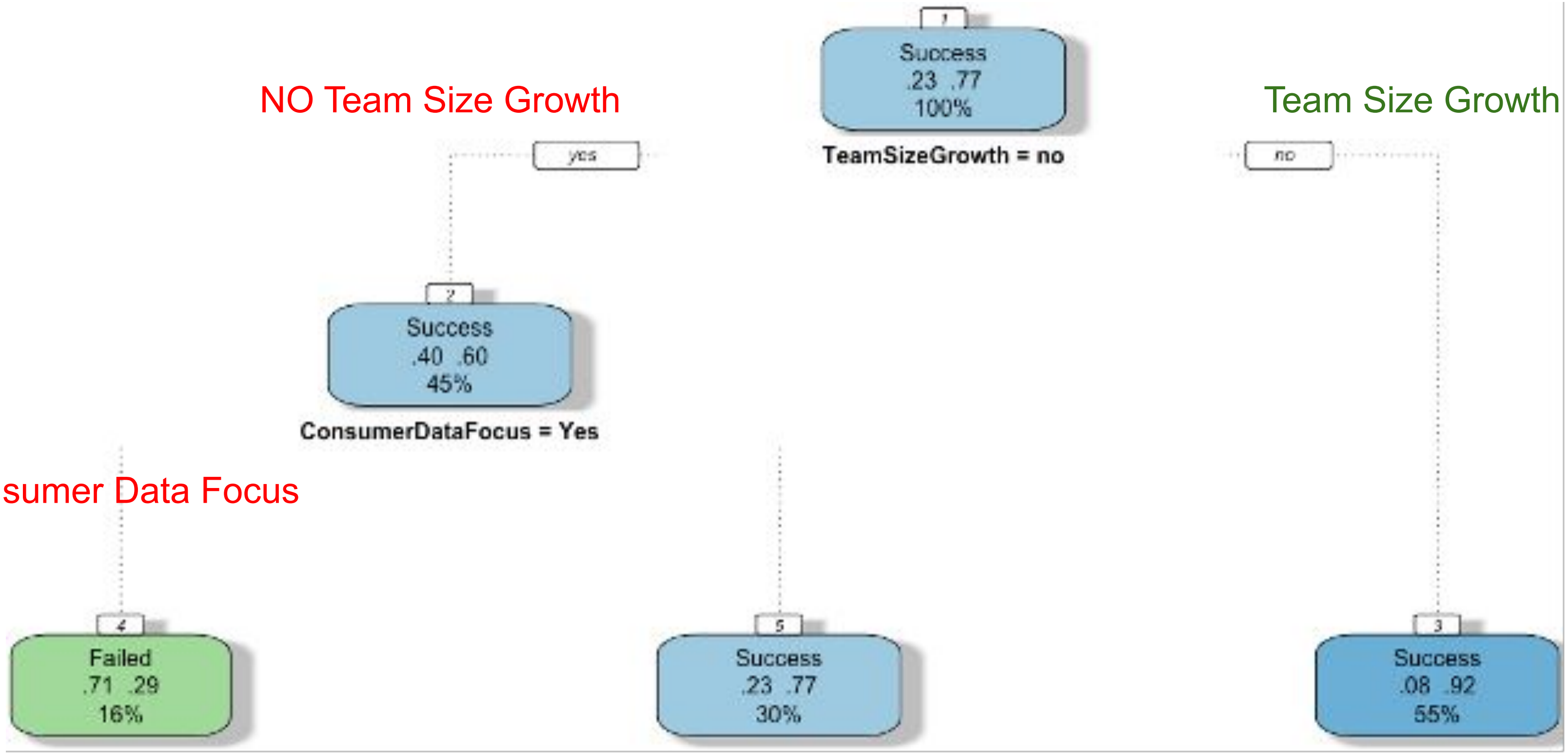
```
> NB_wrong
[1] 8
> NB_error_rate<-NB_wrong/length(category_all)
> NB_error_rate
[1] 0.1311475
```

# Classification and Regression Tree

# CART - All Samples



# CART - Recession Samples



NO Team Size Growth

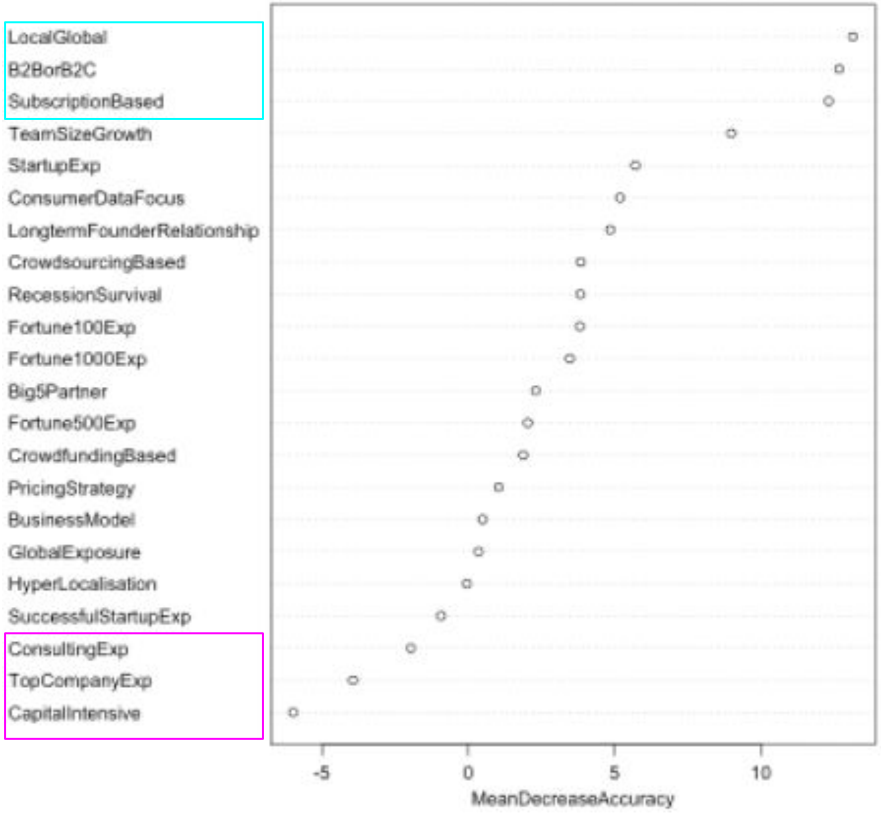
Team Size Growth

Consumer Data Focus

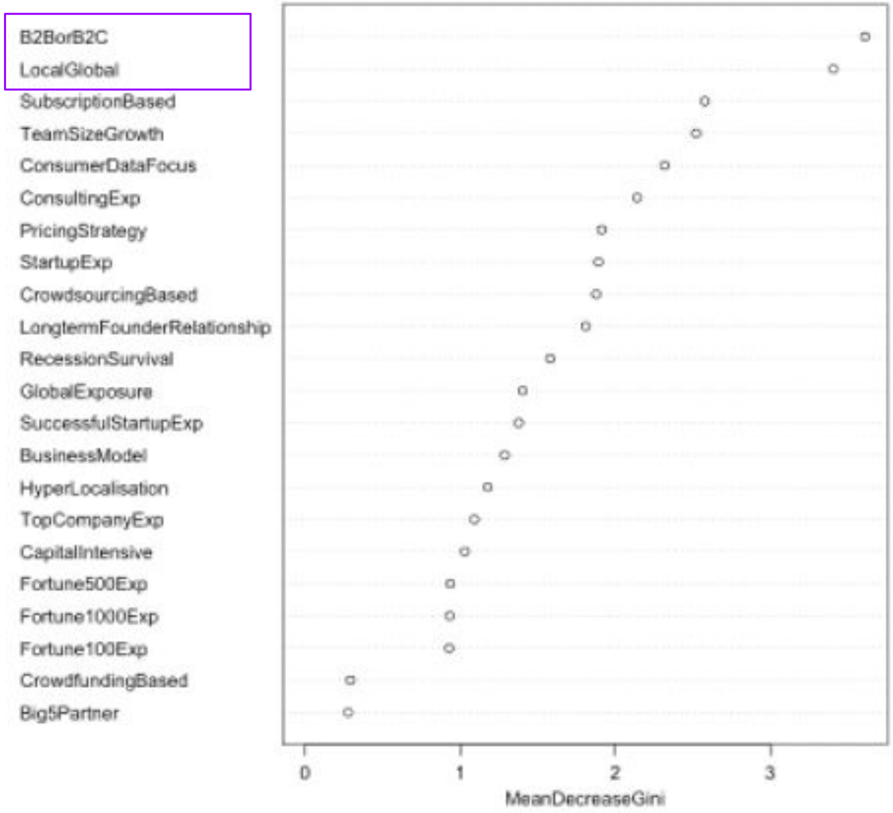
# Random Forest

# Random Forest - All Samples

Mean Decrease Accuracy

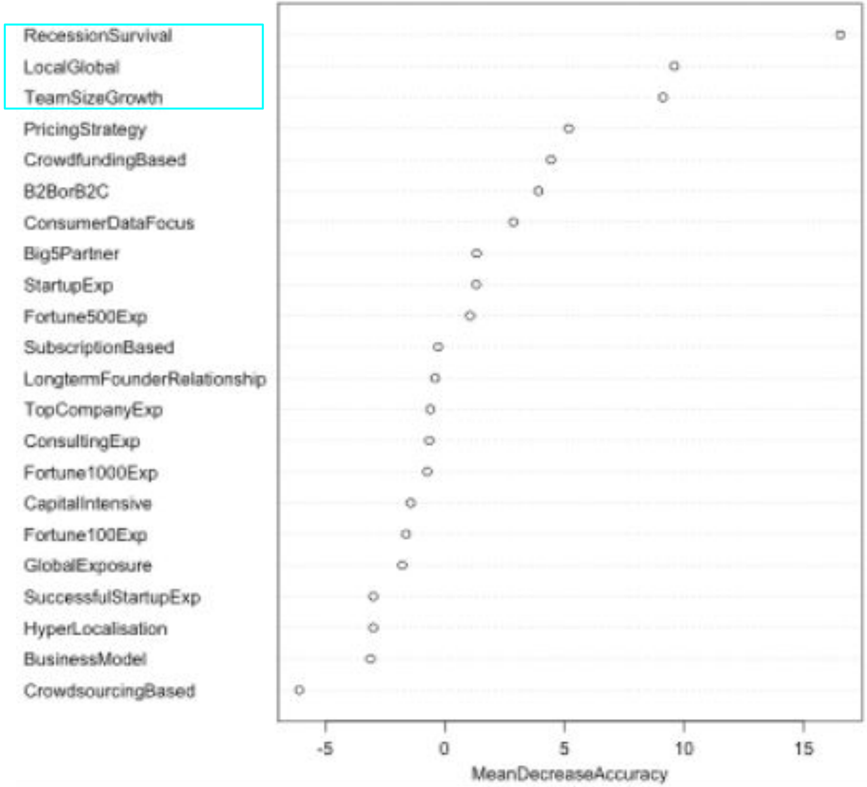


Mean Decrease Gini

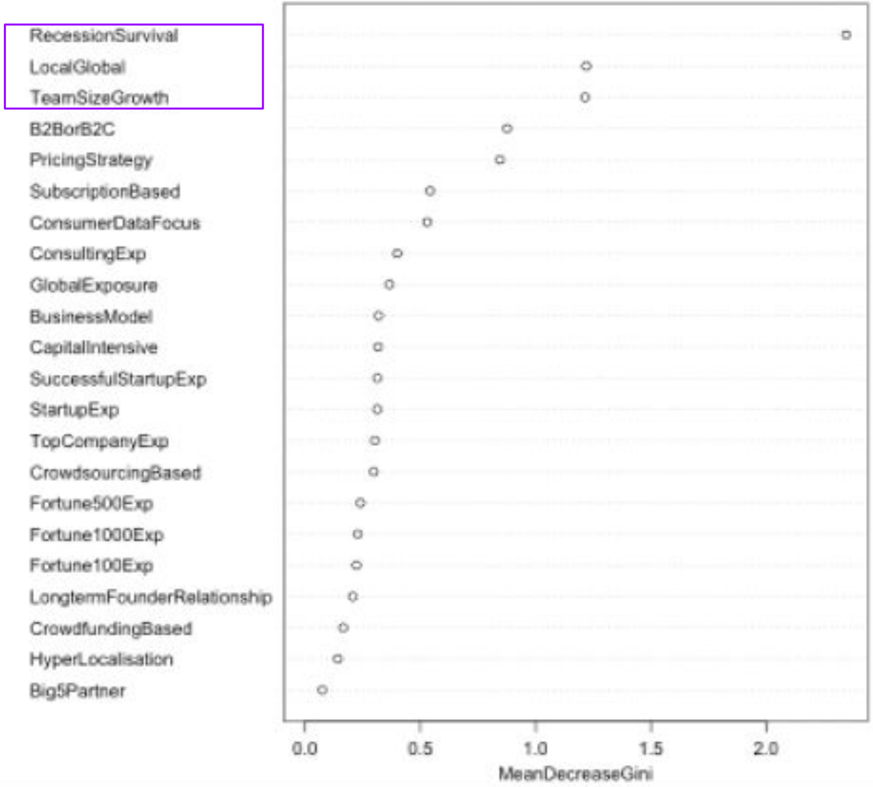


# Random Forest - Recession Samples

Mean Decrease Accuracy

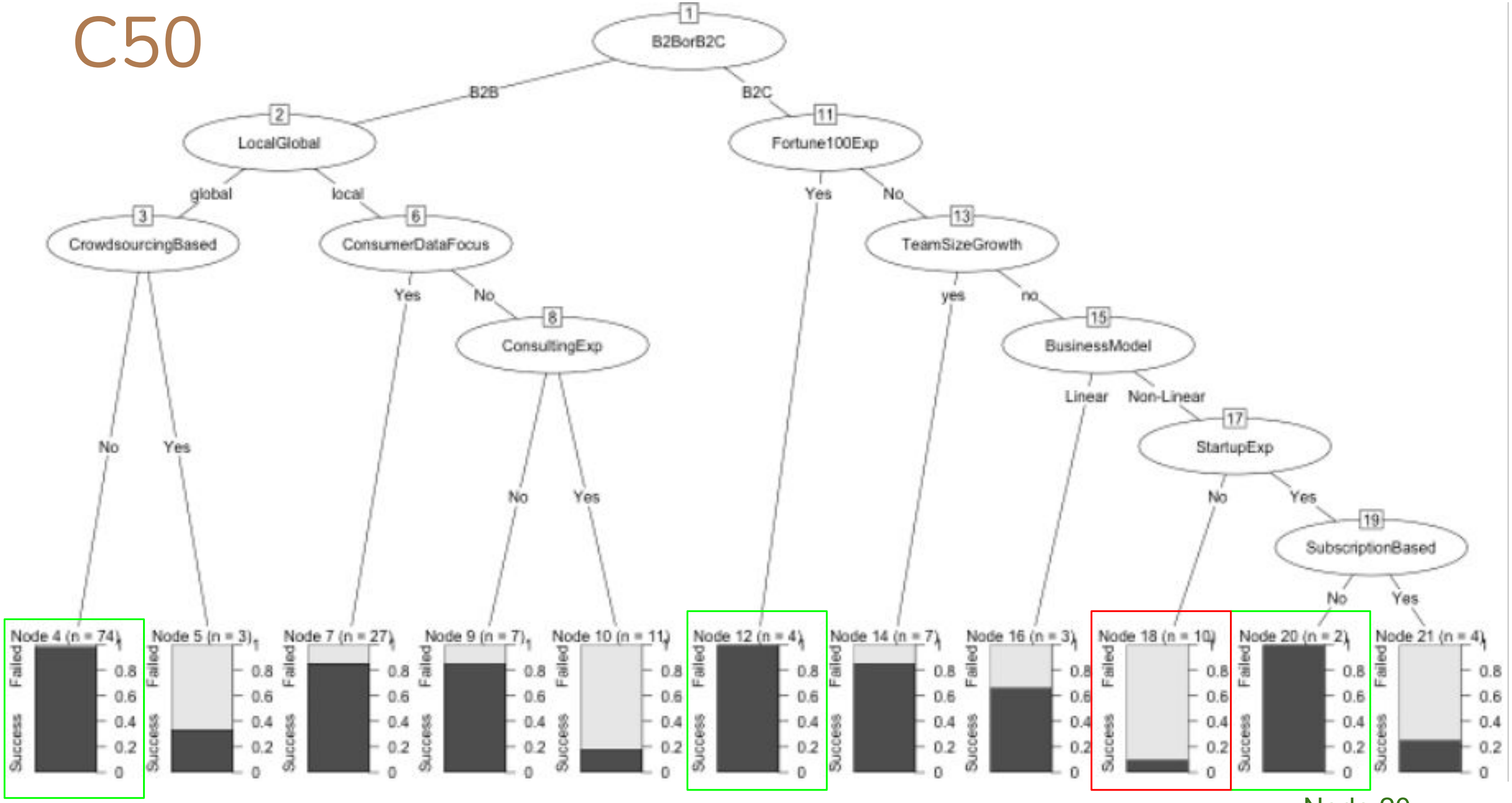


Mean Decrease Gini



# C5.0 Classification

# C50



Node 4

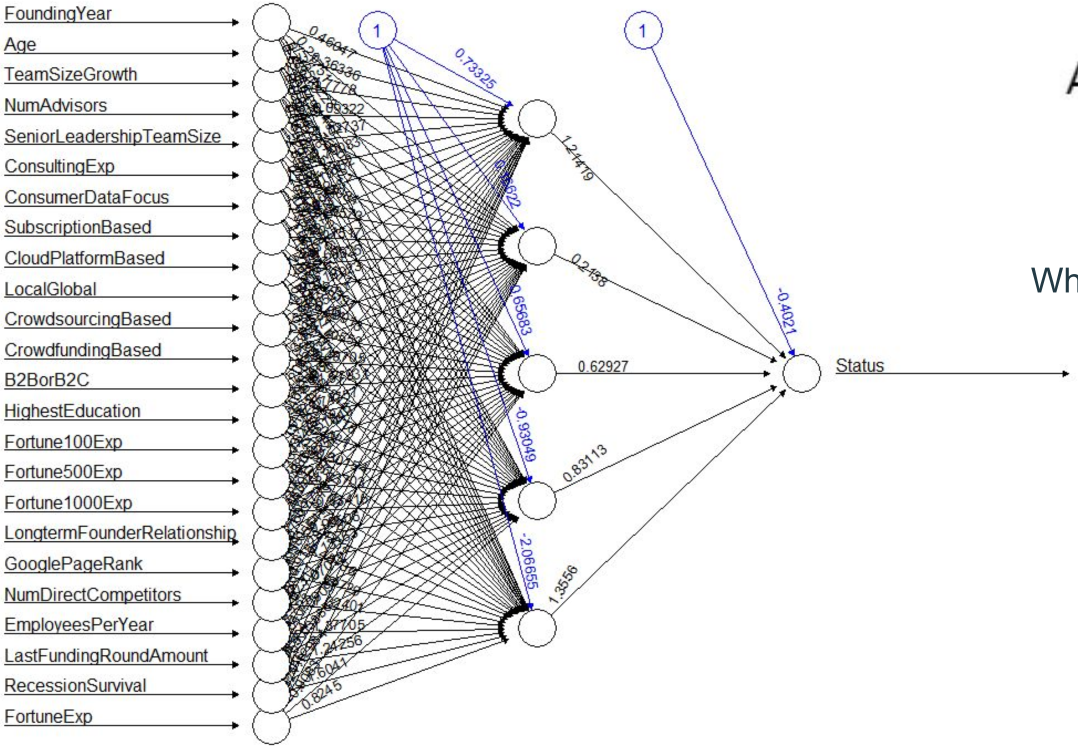
Node 12

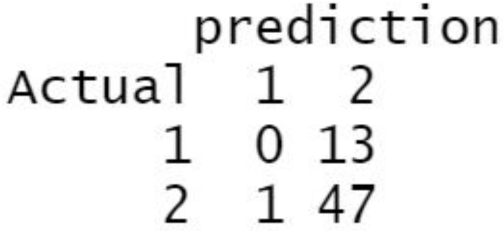
Node 18

Node 20

# Artificial Neural Network Analysis

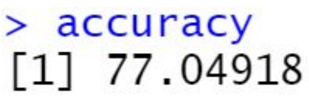
# ANN with 5 Hidden Nodes





```
           prediction
Actual  1  2
     1  0 13
     2  1 47
```

Where 1 = failure and 2 = success



```
> accuracy
[1] 77.04918
```

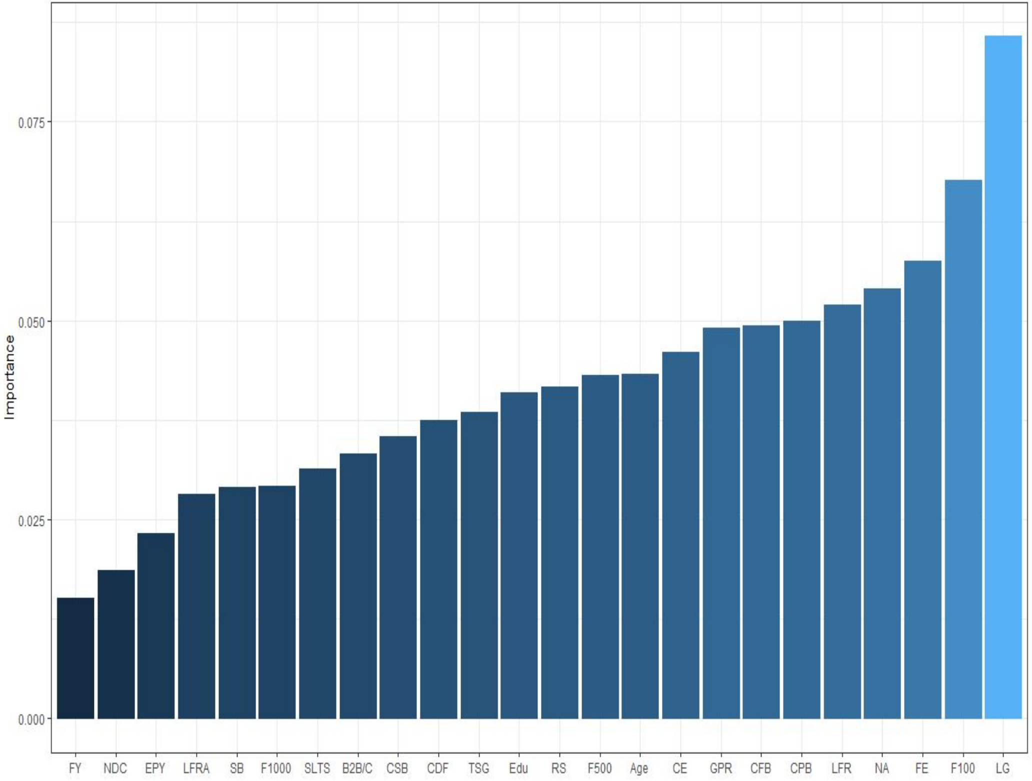
# ANN with 5 Hidden Nodes Analysis

## Acronym list

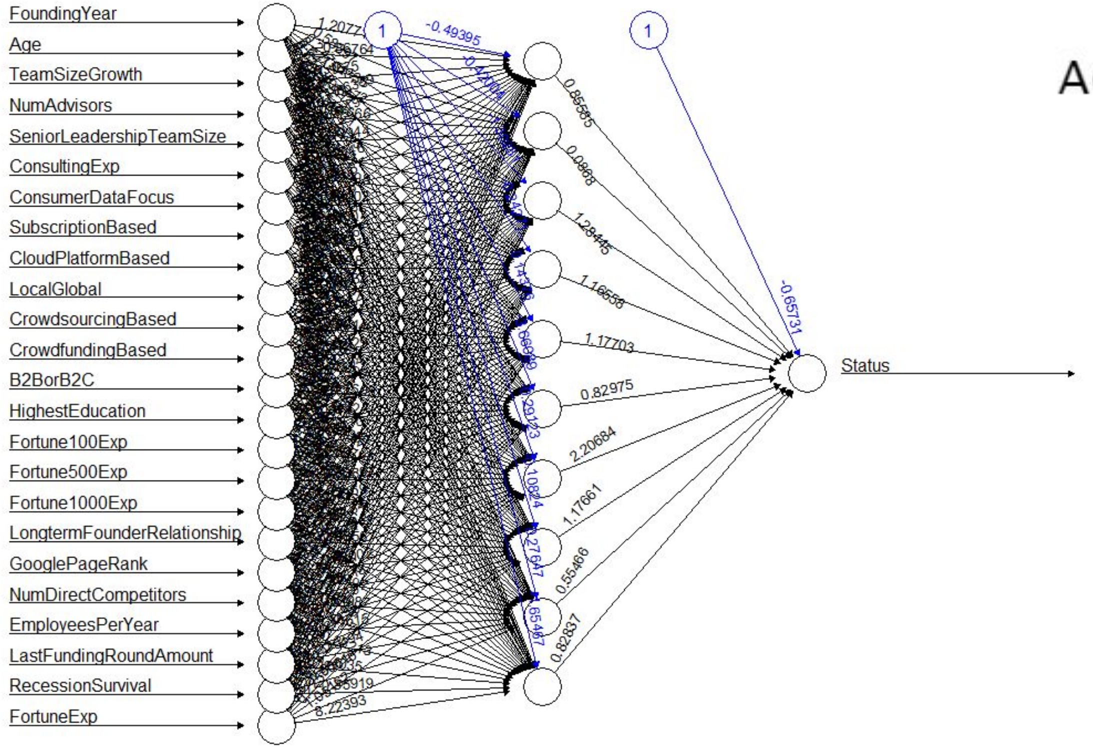
| | | |
|---|---|---|
| FY | = | FoundingYear |
| Age | = | Age |
| TSG | = | TeamSizeGrowth |
| NA | = | NumAdvisors |
| SLTS | = | SeniorLeadershipTeamSize |
| CE | = | ConsultingExp |
| CDF | = | ConsumerDataFocus |
| SB | = | SubscriptionBased |
| CPB | = | CloudPlatformBased |
| LG | = | LocalGlobal |
| CSB | = | CrowdsourcingBased |
| CFB | = | CrowdfundingBased |
| B2B/C | = | B2BorB2C |
| Edu | = | HighestEducation |
| F100 | = | Fortune100Exp |
| F500 | = | Fortune500Exp |
| F1000 | = | Fortune1000Exp |
| LFR | = | LongtermFounderRelationship |
| GPR | = | GooglePageRank |
| NDC | = | NumDirectCompetitors |
| EPY | = | EmployeesPerYear |
| LFRA | = | LastFundingRoundAmount |
| RS | = | RecessionSurvival |
| FE | = | FortuneExp |

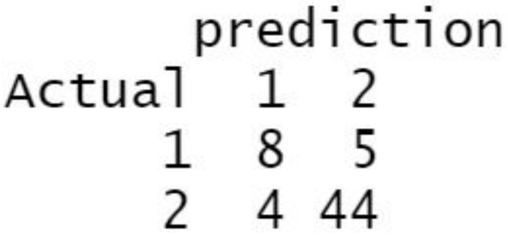
## Importance

```
                                rel_imp
FoundingYear                 0.01512465
Age                          0.04325097
TeamSizeGrowth               0.03844404
NumAdvisors                  0.05396022
SeniorLeadershipTeamSize     0.03145047
ConsultingExp                0.04596629
ConsumerDataFocus            0.03749911
SubscriptionBased            0.02909163
CloudPlatformBased           0.05000867
LocalGlobal                  0.08572718
CrowdsourcingBased           0.03541052
CrowdfundingBased            0.04931589
B2BorB2C                     0.03331078
HighestEducation             0.04097444
Fortune100Exp                0.06763831
Fortune500Exp                0.04309299
Fortune1000Exp               0.02927936
LongtermFounderRelationship  0.05198003
GooglePageRank               0.04915340
NumDirectCompetitors         0.01860393
EmployeesPerYear             0.02333819
LastFundingRoundAmount       0.02812617
RecessionSurvival            0.04173665
FortuneExp                   0.05751609
```



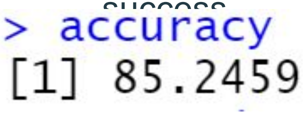
# ANN with 10 Hidden Nodes





```
          prediction
Actual   1  2
     1   8  5
     2   4 44
```

Where 1 = failure and 2 = success



```
> accuracy
[1] 85.2459
```

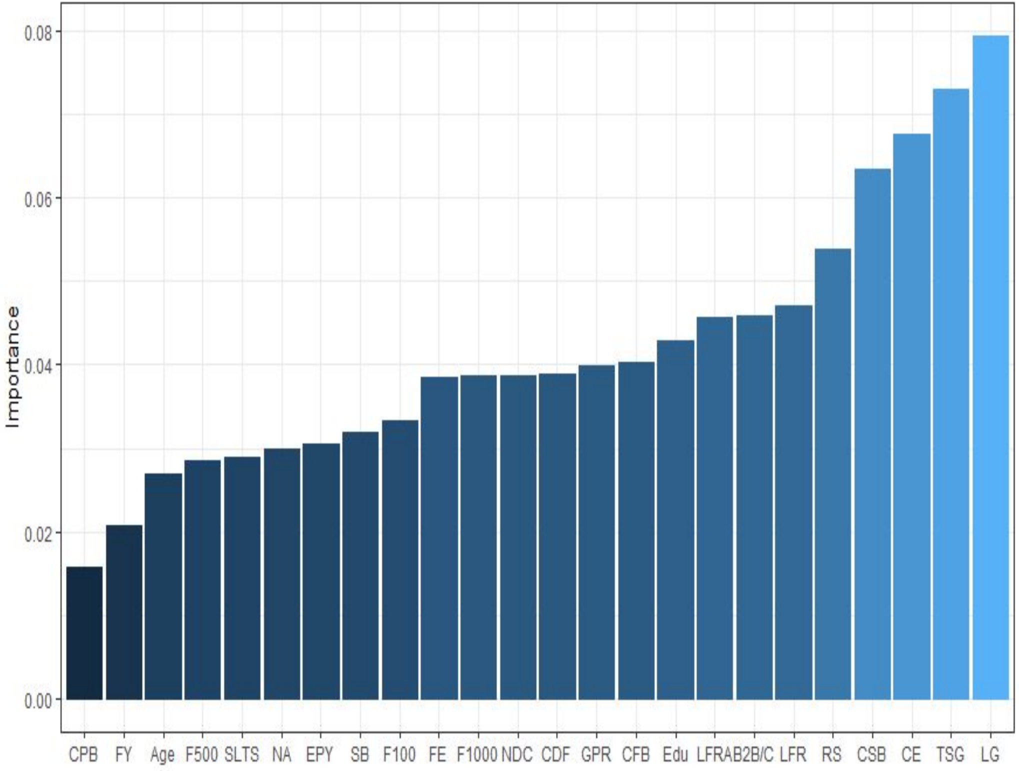
# ANN with 10 Hidden Nodes Analysis

## Acronym list

| | | |
|---|---|---|
| FY | = | FoundingYear |
| Age | = | Age |
| TSG | = | TeamSizeGrowth |
| NA | = | NumAdvisors |
| SLTS | = | SeniorLeadershipTeamSize |
| CE | = | ConsultingExp |
| CDF | = | ConsumerDataFocus |
| SB | = | SubscriptionBased |
| CPB | = | CloudPlatformBased |
| LG | = | LocalGlobal |
| CSB | = | CrowdsourcingBased |
| CFB | = | CrowdfundingBased |
| B2B/C | = | B2BorB2C |
| Edu | = | HighestEducation |
| F100 | = | Fortune100Exp |
| F500 | = | Fortune500Exp |
| F1000 | = | Fortune1000Exp |
| LFR | = | LongtermFounderRelationship |
| GPR | = | GooglePageRank |
| NDC | = | NumDirectCompetitors |
| EPY | = | EmployeesPerYear |
| LFRA | = | LastFundingRoundAmount |
| RS | = | RecessionSurvival |
| FE | = | FortuneExp |

## Importance

```
                                 rel_imp
FoundingYear                  0.02066238
Age                           0.02689374
TeamSizeGrowth                0.07296307
NumAdvisors                   0.03001614
SeniorLeadershipTeamSize      0.02896942
ConsultingExp                 0.06760328
ConsumerDataFocus             0.03881375
SubscriptionBased             0.03187889
CloudPlatformBased            0.01569812
LocalGlobal                   0.07942056
CrowdsourcingBased            0.06332889
CrowdfundingBased             0.04025726
B2BorB2C                      0.04593928
HighestEducation              0.04280510
Fortune100Exp                 0.03334516
Fortune500Exp                 0.02850328
Fortune1000Exp                0.03866388
LongtermFounderRelationship   0.04704763
GooglePageRank                0.03990291
NumDirectCompetitors          0.03868772
EmployeesPerYear              0.03054412
LastFundingRoundAmount        0.04569248
RecessionSurvival             0.05386211
FortuneExp                    0.03850082
```



# 4.) ACQUIRE WISDOM TO UNDERSTAND OPTIMAL ACTIONS

# Key Takeaways

**Business Models:**

The Business-to-Business model has an advantage over Business-to-Consumer model.

Global startups have a significant advantage over local startups.

**Properties of Employees**:

Quality of employees has a greater impact on the success of a business during a recession.

Consulting experience usually has a negative impact on success.

Startup Experience has a positive impact regardless of previous failure.

**Future Research:**

To prevent overfitting or underfitting a future model, all latter models should aim for a minimum of 70% kNN and Naive Bayes accuracy.

# Future Research

- Business to Business Startups:
  - Question: Why is Business to Business model more effective than Business to Consumer?
  - Hypothesis: Easier to get market (less sales, easy to establish trust, less competition)
- Local vs Global Startups:
  - Question: Why do global startups have a significantly higher chance of success?
  - Hypothesis: most less experienced groups go local (less cost, less expertise needed), while more experienced groups go global (higher monetary gain, significant expertise needed)

# Future Research (Continued)

- Start-Ups by Consultants
  - Question: Why does consulting experience negatively impact success?
  - Hypothesis: Consultants have expertise on specific company functions while neglecting overall company functionality which may lead to decreased start-up success.
- Start-Ups during Economic Recession
  - Our dataset only had 59 recession samples, so another future study with more samples allow for classification models of higher accuracy.

# Bibliography

Kaggle: https://www.kaggle.com/sujithsherigar/startup-success-rate-analysis?select=CAX_Startup_Data.csv

Neural Net Tools package: https://rpubs.com/julianhatwell/annr

Questions?