INDIAN STARTUP FUNDING

Question: What role do time and location play in funding received by startups

63.76% startups located in **4** megacities

$14.5 Billion funded in **2019**

Goal: Picture how startups receive funding based on time and location

Software: R, Python

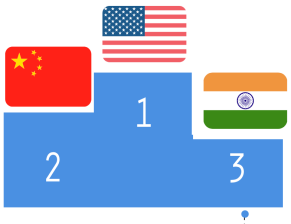
Models: Decision Tree, Logistic regression model

Tools: Pictochart, Excel

Data: Gathered by Trak.in, dataset compiled on Kaggle.com

By Anushree Soman
ID: 23892873

Motivation for Analysis



#3
Globally
Number of startups

15%
New startups yearly in need of funds

108%
Growth in funding since 2017

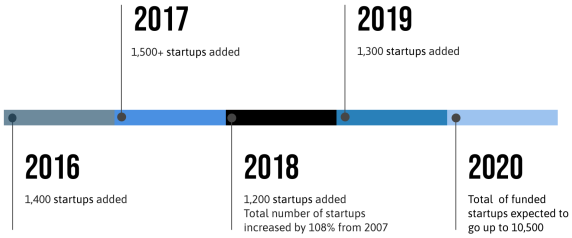
Why Location is important?

Potential Investors: Investors in cities prefer familiar markets.

Talent pool: Easier to recruit new highly qualified employees .

Business Friendly: Governments offer tax benefits and other incentives to businesses in cities

Scalability: Helps retail, and transport business scale due to high demand for their service.



Why Time is important?

Market conditions: Market conditions change fast causing many startups to fail.

Competitors: If not executed on time another competitor might market a similar concept

If Uber started in 1980's it might not have done well due to lack of demand.

Summary

Basic Dataset statistics

2459 | Total startups in the dataset

68.1% — More than 2/3rd of startups listed have information on the exact amount they received.

31.9% — About 1/3rd of startups listed were either not funded, or did not provide the information needed.

By Location

In this dataset, 63.8% of startups were based in 4 Megacities

They are Bangalore, Mumbai, New Delhi, and Chennai.

Nearly 23% were based in just 1 city, Banglore.

23%

By Industry

31%

In the dataset, 31% of startups operated in the Consumer Internet industry.

Nearly 15.7% of startups operated in the Technology industry, the second-highest.

15.7%

By Investment Type

45%

In the dataset, 45% of startups were funded by Private Equity and 45% by Seed Funding or a combination of Angel and seed investment.

Nearly 90% of all investment came from these two sources.

45%

By Amount Received

23.5%

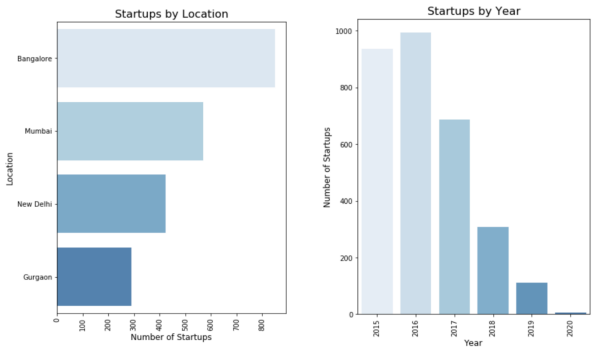
More than 2/3rd were funded some amount.

However, only 2.4% got more than 100million.

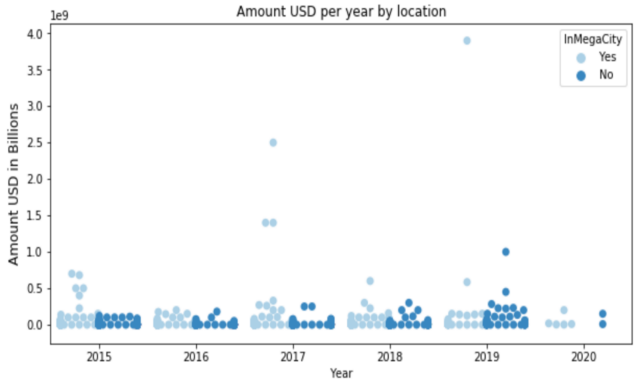
2.4%

Basic Statistic Graphs

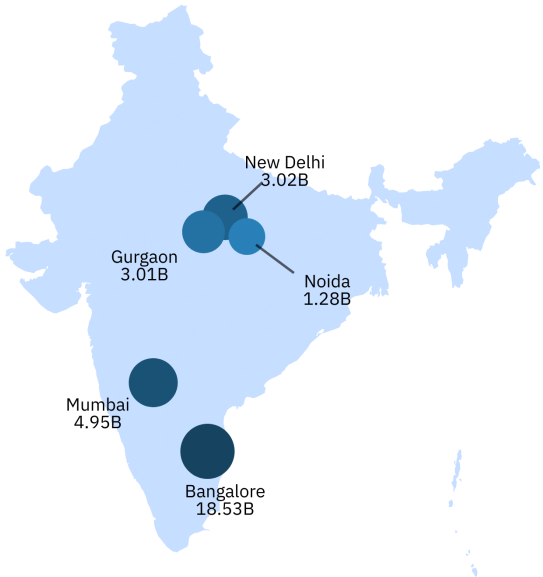
By Location and Year



- **850** of 2459 startups based in Bangalore
- Mumbai, New Delhi, and Gurgaon follow at 570, 424, and 291 respectively.
- Maximum startups were funded in **2016**.
- However, this does not mean the growth has slowed as the dataset could have missing values for other years.



The amount received by startups has increased each year. The year 2019 saw the single maximum amount invested to date, at 3.9Billion

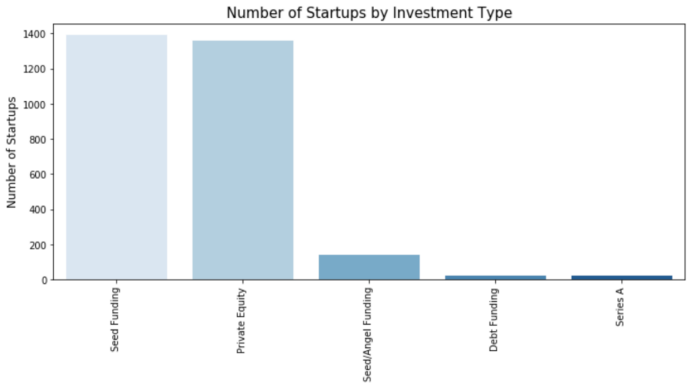


Map showing top 5 cities by total amount invested in Billion USD 2015-2020

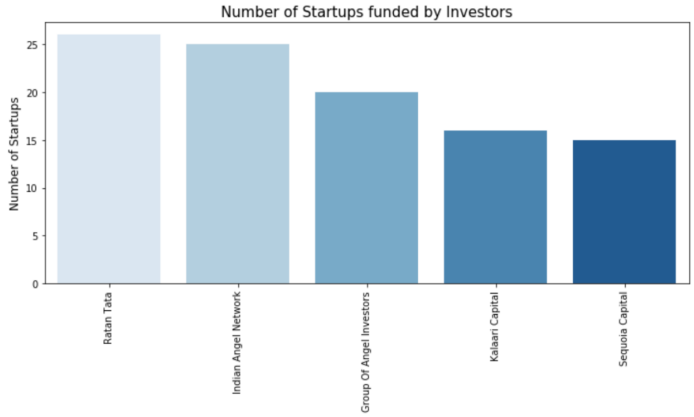
Basic Statistic Graphs

Other variables

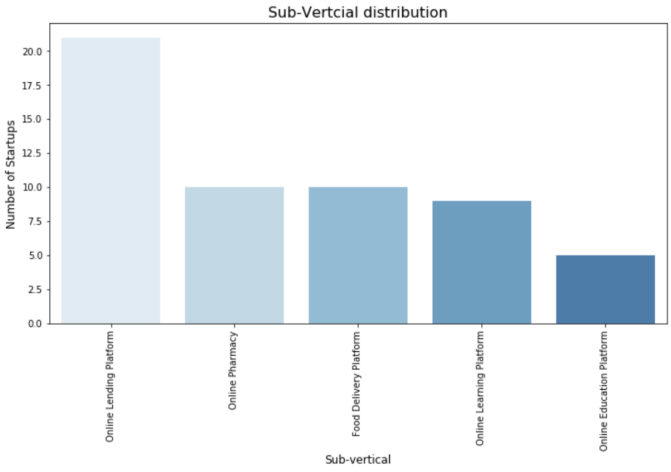
Seed Funding and Private Equity are the most common Investment Types for startups



Ratan Tata, the top known investor has funded more than 25 startups.



More than 20 startups are operating in the Online Lending platform sub-vertical followed by Online Pharmacy.

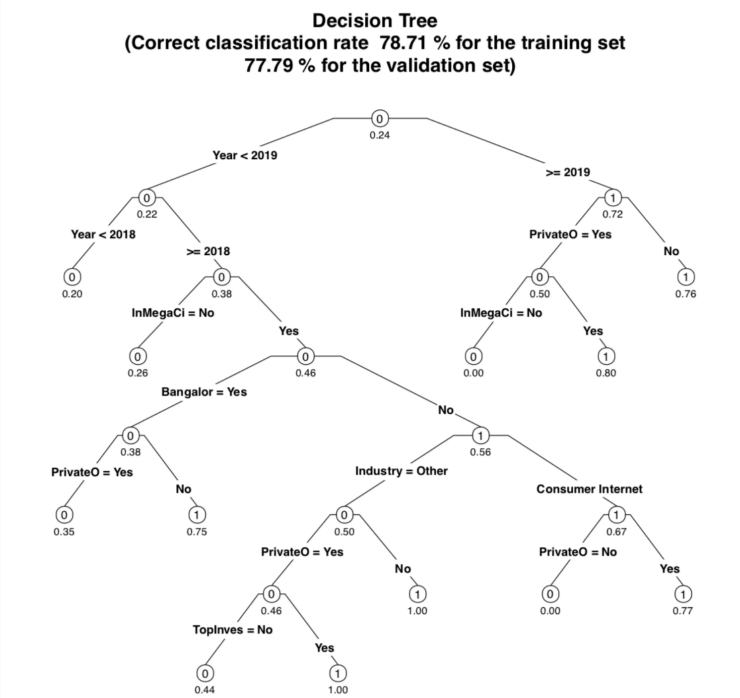


The Statistical Models

1
Minsplit

Decision Tree Using R

.0001
Complexity Factor



rpart(formula = FiveMillionOrMore ~ Industry + InMegaCity + BangaloreBased + TechRelated + PrivateOrSeed + Year + TopInvestor, data = trainingSet, method = "class", control = rpart.control(minsplit = 1, cp = 1e-04))

Calculated variables

Variable	Description
Year	The year in which startup was funded
InMegaCity	The startup CityLocation is Bangalore, Delhi, Mumbai or Chennai
BangaloreBased	The startup CityLocation is Bangalore
PrivateOrSeed	InvestmentType is by Private Equity or Seed Funding
TopInvestor	InvestorName is Ratan Tata or Undisclosed Investors
Industry	Whether IndustryVertical is Consumer Internet or Other

Confusion Matrix

OBSERVED / PREDICTED	0	1
0	1,126	297
1	27	72

Training Set

Validation Set

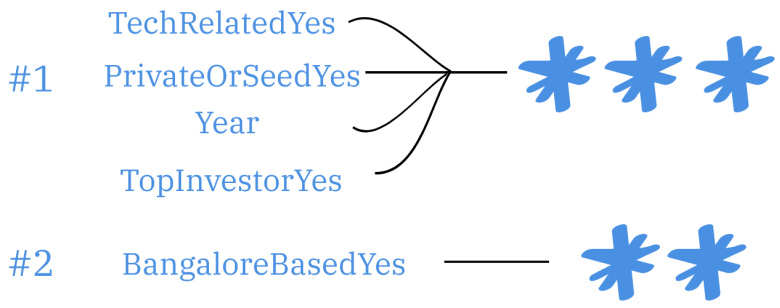
OBSERVED / PREDICTED	0	1
0	1,132	295
1	43	52

The Statistical Models

Logistic Regression Model

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -618.20967   95.82579  -6.451 1.11e-10 ***
BangaloreBasedYes    0.30995    0.11089   2.795 0.005189 **
IndustryOther        0.18937    0.13556   1.397 0.162415
TechRelatedYes      -0.48043    0.12830  -3.745 0.000181 ***
InMegaCityYes       -0.01166    0.10704  -0.109 0.913263
PrivateOrSeedYes    -0.76821    0.21084  -3.644 0.000269 ***
Year                 0.30640    0.04749   6.453 1.10e-10 ***
TopInvestorYes      -1.73002    0.36791  -4.702 2.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Significant Variables



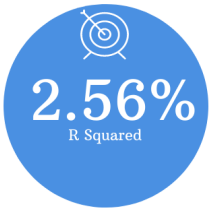
Interpretation of coefficients

- Industry and Year have a positive relationship with getting more than Five million USD.
- Whether Tech Related, in Mega City, Funding Type, and Top Investor have an adverse relationship with funding amount.

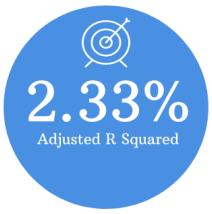
Summary

Industry, Year, Investor, Investment Type, and Location are important factors deciding amount funded

The Statistical Models



Linear Regression Model (LRM)



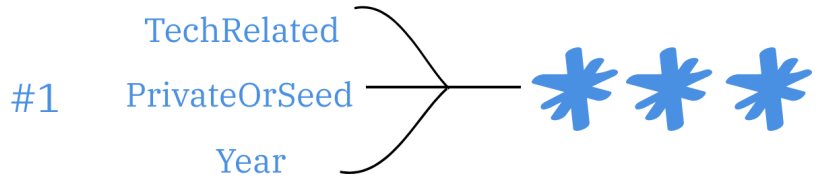

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.584e+10  4.231e+09  -3.744 0.000184
IndustryIsConsumerInternet  3.493e+06  5.414e+06   0.645 0.518765
BangaloreBased              1.324e+07  4.561e+06   2.904 0.003714
InMegaCity                  1.928e+05  4.249e+06   0.045 0.963810
TechRelated                -1.819e+07  5.445e+06  -3.340 0.000847
PrivateOrSeed              -3.629e+07  9.927e+06  -3.655 0.000261
TopInvestor                -6.389e+06  7.939e+06  -0.805 0.421003
Year                        7.882e+06  2.096e+06   3.760 0.000173

(Intercept)                ***
IndustryIsConsumerInternet
BangaloreBased             **
InMegaCity
TechRelated                ***
PrivateOrSeed              ***
TopInvestor
Year                       ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99150000 on 3036 degrees of freedom
Multiple R-squared:  0.02555,   Adjusted R-squared:  0.02331
F-statistic: 11.37 on 7 and 3036 DF,  p-value: 2.594e-14
```

Significant Variables

- When the p-value is smaller than .05, variables can be considered as significant.



Summary

TechRelated, Year, Investment Type, and Location are important factors deciding amount funded

Model Evaluation

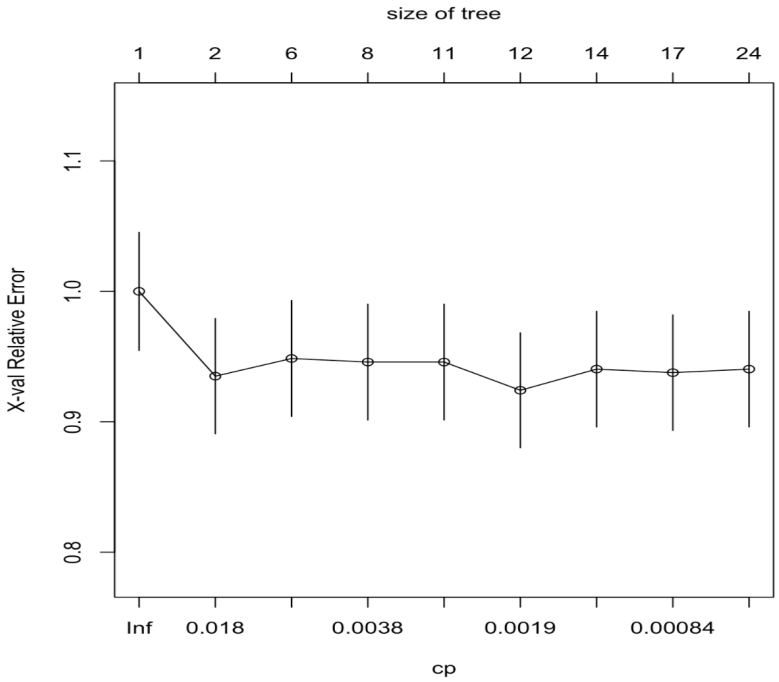
Decision Tree

78.71%
Training Set

Accuracy

77.79%
Validation Set

- Correct prediction of whether a startup will receive more than Five million for 78.71% of Training set values and 77.79% of validation set values.
- Model does not seem to have an overfitting issue as accuracy is less than 80% for both sets.
- The lower minsplit value makes the decision tree model better and more complex as adding more splits requires a lesser number of minimum observations.
- The smaller complexity factor makes our tree more complex and accurate.



Observations

- 12 leaf nodes, nodes with no children
- Year is the most significant variable to split according to
- Before 2019, second most important factor to split by was the location if InMegaCity
- After 2019, second most important factor to split by is PrivateOrSeed Funding
- Location if InMegaCity followed as the third most important variable.

Model Evaluation

Logistic Regression

Model

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5792  -0.7441  -0.6518  -0.2851   2.5130

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -618.20967   95.82579  -6.451 1.11e-10 ***
BangaloreBasedYes    0.30995    0.11089   2.795 0.005189 **
IndustryOther        0.18937    0.13556   1.397 0.162415
TechRelatedYes      -0.48043    0.12830  -3.745 0.000181 ***
InMegaCityYes       -0.01166    0.10704  -0.109 0.913263
PrivateOrSeedYes    -0.76821    0.21084  -3.644 0.000269 ***
Year                 0.30640    0.04749   6.453 1.10e-10 ***
TopInvestorYes      -1.73002    0.36791  -4.702 2.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3321.0  on 3043  degrees of freedom
Residual deviance: 3122.4  on 3036  degrees of freedom
AIC: 3138.4

Number of Fisher Scoring iterations: 5
```

Observations

- 3321.0 on 3043 DoF, the null deviance shows how well the amount funded is predicted by a model including only the intercept.
- Including the independent variables does not reduce the deviation by much, 3122.4 on 3036 DoF
- Fisher’s Scoring Algorithm needed 5 iterations to perform the fit.
- The Deviance residual is how much the model got each prediction wrong or how different the predictions were from the actual results.
- The dataset seems to have a large enough sample size to perform the regression

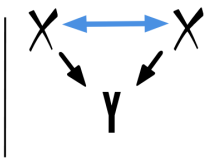
Assumptions

Ommited variables

- Possibility of omitted variables.
- The location of investors can affect AmountUSD funded and also which IndustryVertical is preferred by them.

Lack of multicollinearity

- 2 or more independent variables should not be highly correlated with each other
- This dataset does not seem to have such problem



Summary

We cannot say the model is perfect, however it is good for showing the regression between regressors and y variable value

Model Evaluation

Linear Regression Model

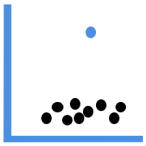
Assumptions

Exogeniety vs Endogeniety

- **Ommited variables**: Possibility of omitted variables. Location of investors can affect AmountUSD funded and also which IndustryVertical is preferred by them.
- **Measurement errors**: Possible reporting errors, due to some companies not providing enough or accurate information about the topic.
- **Simultaneity**: Possible issue due to AmountUSD(y) determined by city location(x) and vice versa where more funds can enable opening business in city.

Randomness of sample

- No information about the randomness of the sample.
- But only startups that were funded are included in the dataset.
- Randomness of data might be a problem.

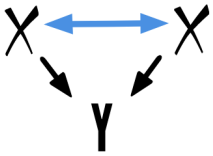


Outliers in data

- There is one outlier in the data for AmountUSD(y)
- The amount was funded to Rapido Bike.
- Examination of outlier shows data is good and thus not removed from the model for this dataset.

Lack of multicollinearity

- 2 or more independent variables should not be highly correlated with each other
- This dataset does not seem to have such problem



Overall

- The models chosen seem good, but violate a few assumptions of regression
- For LRM, R squared and Adjusted r squared is very low at 2%
- This is the measure of how close the data is fitted to the line
- However, a lower R square doesn't mean the model is incorrect
- An omitted variable bias can cause a lower value.

Summary

We cannot say the model is perfect, however it is good for showing the regression between regressors and y variable value

Model interpretation

Implications

Top funding Factors

#1 Year the startup was funded

#2 Private Equity/Seed Fund

#3 Location of the startup

#4 Industry Vertical of startup

Recommendations

Research Market Conditions

- Conduct thorough research for market conditions before starting a business and looking for funding.
- This can improve the chance to get higher funding.

Time your actions

- Startups need to calculate and manage their time to ensure they enter the market at the correct opportunity.
- This is good as it gives all startups equal opportunity to grow.

Choose Location wisely

- Conduct thorough market research for many cities before choosing a location.
- Base the startup in megacities as these attract the most funds.
- This can improve the chance to get higher funding.

Target funders by Investment type

- Private Equity and Seed Funding Investment Types fund the maximum startups
- While looking for investment target companies that provide seed funds or private equity.