Indian Startup Growth Analysis

Name: Akarsha Mandrekar

Date: 10th December 2021
(Project-1)

~

"Errors using inadequate data are much less than those using no data at all."

~ Charles Babbage ~
(Inventor & Mathematician)

~

Abstract:

India is a developing country and there are many potential start-ups that are emerging in this vast market with population of over 1.39 billion. That's definitely a huge number and a bigger market, that naturally gives us an boost to have a own venture, but without proper analysis and with low understanding about the nature of the Indian market one can have a huge downfall. It is very important to look at all the parameters before taking this important decision. More than 100 million start-ups are launched per year, which is about 3 start-ups per second. But more than 50% of start-ups fail in the initial four years. There are various reasons for a start-up to fail for example lack of focus, raising too much money too soon, lack of general and domain-specific business knowledge, etc.

Lets look at some Advantages and disadvantages of start-ups.

Important Factors:

- Investments and Funding's
- Location feasibility of start-up
- Education of CORE committee members
- Number of Milestones achieved parodically
- University of Core committee
- Domain

1.Problem Statement:

The algorithm proposed in this paper will help to predict the success of a start-up company. In this test model we will take some of these important parameters to train our model and build a system that will help the existing start-ups, investors as well as the new-developing start-ups. This paper will aid start-up companies to know which factors are essential for getting an investment. The algorithm will be based on more than 3500 companies' data collected from Kaggle.com. A variety of methods can be used to determine the best model such as random forest, text parsing, logistic regression, decision tree and survival analysis.

"Big data is at the foundation of all the mega trends that are happening, from social to mobile to cloud to gaming"

~Chris Lynch

2. Business Need Assessment:

In this assessment we will find out what does start-up environment needs the most for it to flourish. The model targets the area that prevents the startups from reaching its desired goals and bridges the gap between R&D and implementation as per the start-up requirement.

Data analysis is very important tool in a startup as it helps determine the right starting point for any project, investment, innovation and optimization. With the help of different data present of successful and of growing ventures well will design a model to predict the success of any firm.

- Data with details of various start-ups and companies.
- Sorting the parameters that affect the start-up growth.
- Algorithm to understand the trend.

- Age of the start-ups.
- Location of stratus.
- Funding's gained by start-ups.

3.Target Specifications and Characterization:

The market for environmentally-friendly investments is rapidly growing. Unfortunately, many of the wealthiest investors are often reluctant to invest in start-ups – especially impact start-ups – due to the lack of relevant research available, the higher risks involved, and the lack of business history that can predict future success. Investing in start-ups is globally understood to be risky and not an optimal asset management strategy. However, there are plenty of start-ups with tremendous potential, but they may lose out on funding and success due to this bias and lack of insight, trust, and understanding of start-up investing. If a project doesn't already have funding, it becomes difficult to raise new funding.

4. External Search (Information sources/References):

This dataset we used has funding information of the Indian start-ups from January 2015 to November 2021. It includes columns with the date funded, the city, the start-up is based out of, the names of the funders, and the amount invested (in USD).

- Dataset Origin:
  https://trak.in/india-startup-funding-investment-2015/

- Kaggle Dataset:
  https://www.kaggle.com/sudalairajkumar/indian-startup-funding/

References:

- Failed Start-ups In India (Case Study)
  https://startuptalky.com/why-startups-fail-case-study/

- Which factors Determine the Success or failure of startup companies?
  https://www.grin.com/document/372343

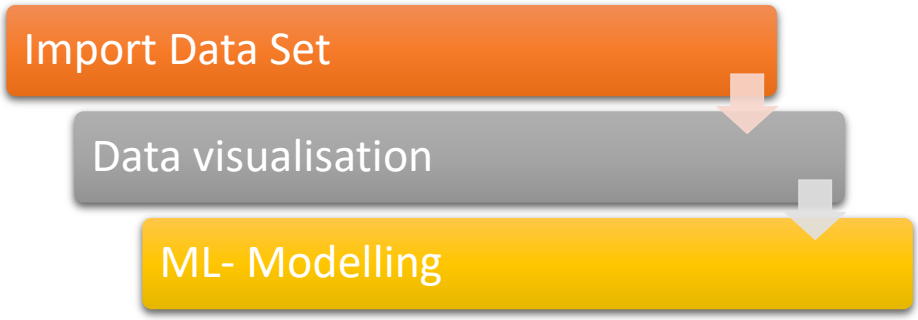
5. Applicable Constraints:

1. Regular Collection and Updating the dataset.
2. Data Collection breaking the protocols of Privacy Regulations Of the start-ups.
3. Tedious task for cleaning the data.
4. The use of cloud platforms to store the data gathered over the net.

6. Business Opportunities:

The target customers here are mainly small start-ups and people who are looking forward to do investments in the small scale start-ups. It will also help small start-ups to meet the requirements for a successful venture. If the model is successfully implemented for small start-ups then the model can be expanded for unicorn start-ups.

7. Concept Generation:



Data was imported from Kaggle and the parameters considered for developing this model were Age of the start-up i.e. when it was incorporated, The funding received by each start-up, name and type of the investors, and the origin of the start-ups. We also took into consideration the Domain type of each Industry to analyse which sector receives most funding.

To make this model more effective we can also include various other parameters like :

- Qualification level of the core team members.
- Alma mater of the Team Members.
- Burn Rate of the venture.
- Revenue generated by the company.
- The millstones achieved.

9. Code Implementation:

1) Loading and Cleaning the Dataset

- Importing the libraries

In [1]:
```
import numpy as np                #to perform a variety of arithmetic operations on array
import pandas as pd               #for quantitative analysis
import matplotlib.pyplot as plt   #plotting
import seaborn as sns             #comes under matplotllib, for visualizing random distribution
color=sns.color_palette()

import plotly.offline as py           #to initialise plots inside a notebook
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go
pd.options.mode.chained_assignment=None
```

In [2]:
```
fd=pd.read_csv("startup_funding.csv")
fd.head()
```

Out[2]:

	Sr No	Date dd/mm/yyyy	Startup Name	Industry Vertical	SubVertical	City Location	Investors Name	InvestmentnType	Amount in USD	Remarks
0	1	9/1/2020	BYJU$\diamond$S	E-Tech	E-learning	Bengaluru	Tiger Global Management	Private Equity Round	20,00,00,000	NaN
1	2	13/01/2020	Shuttl	Transportation	App based shuttle service	Gurgaon	Susquehanna Growth Equity	Series C	80,48,394	NaN
2	3	9/1/2020	Mamaearth	E-commerce	Retailer of baby and toddler products	Bengaluru	Sequoia Capital India	Series B	1,83,58,860	NaN
3	4	2/1/2020	https://www.wealthbucket.in/	FinTech	Online Investment	New Delhi	Vinod Khatumal	Pre-series A	30,00,000	NaN
4	5	2/1/2020	Fashor	Fashion and Apparel	Embroiled Clothes For Women	Mumbai	Sprout Venture Partners	Seed Round	18,00,000	NaN

- Data cleaning

In [8]:
```
#fd.loc[20,"Amount in USD"]="4,20,00,000"
#fd.loc[89,"Amount in USD"]="4,20,00,000"
#fd.loc[91,"Amount in USD"]="4,20,00,000"
#fd.loc[fd["Amount in USD"]=="undisclosed"]
```

In [9]:
```
#since "Amount in USD" is a str we need to convert it else numeric operations wont work
fd["Amount in USD"] = fd["Amount in USD"].apply(lambda x: float(str(x).replace(",","")))
fd["Amount in USD"] = pd.to_numeric(fd["Amount in USD"])
fd.head()
```

Out[9]:

	Sr No	Date dd/mm/yyyy	Startup Name	Industry Vertical	SubVertical	City Location	Investors Name	InvestmentnType	Amount in USD
0	1	9/1/2020	BYJU$\diamond$S	E-Tech	E-learning	Bengaluru	Tiger Global Management	Private Equity Round	200000000.0
1	2	13/01/2020	Shuttl	Transportation	App based shuttle service	Gurgaon	Susquehanna Growth Equity	Series C	8048394.0
2	3	9/1/2020	Mamaearth	E-commerce	Retailer of baby and toddler products	Bengaluru	Sequoia Capital India	Series B	18358860.0
3	4	2/1/2020	https://www.wealthbucket.in/	FinTech	Online Investment	New Delhi	Vinod Khatumal	Pre-series A	3000000.0
4	5	2/1/2020	Fashor	Fashion and Apparel	Embroiled Clothes For Women	Mumbai	Sprout Venture Partners	Seed Round	1800000.0

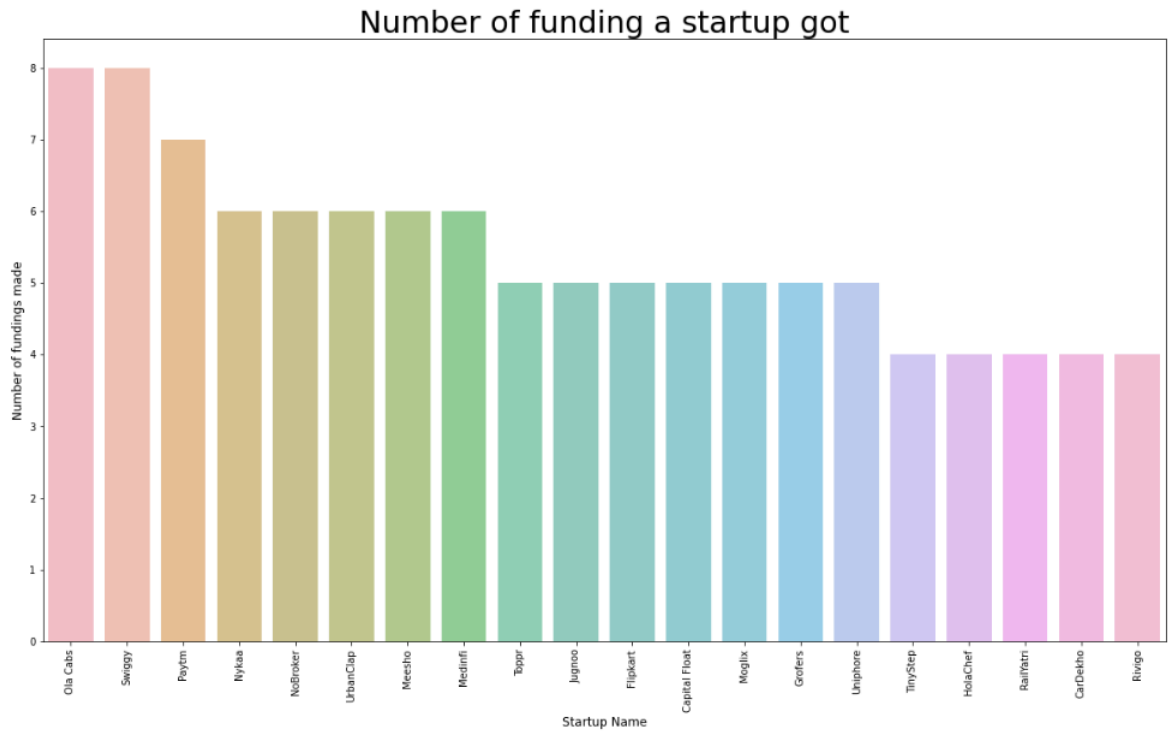
In [10]:
```
#now we will fix the Date cloumn and write proper format
fd['Date dd/mm/yyyy'][fd['Date dd/mm/yyyy']=='12/05.2015'] = '12/05/2015'
fd['Date dd/mm/yyyy'][fd['Date dd/mm/yyyy']=='13/04.2015'] = '13/04/2015'
fd['Date dd/mm/yyyy'][fd['Date dd/mm/yyyy']=='15/01.2015'] = '15/01/2015'
fd['Date dd/mm/yyyy'][fd['Date dd/mm/yyyy']=='22/01//2015'] = '22/01/2015'
```

2) Data visualisation

- Number of Funding's received

In [16]:
```
#compnay who got maximum number funding
print("Total startups funded : ", len(fd["Startup Name"].unique()))
print(fd["Startup Name"].value_counts().head(10))
startupname = fd['Startup Name'].value_counts().head(20)
plt.figure(figsize=(20,11))
sns.barplot(startupname.index, startupname.values, alpha=0.6)
plt.xticks(rotation='vertical')
plt.xlabel('Startup Name', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Number of funding a startup got", fontsize=30)
plt.show()
```

```
Total startups funded :  2457
Ola Cabs     8
Swiggy       8
Paytm        7
Nykaa        6
NoBroker     6
UrbanClap    6
Meesho       6
Medinfi      6
Toppr        5
Jugnoo       5
Name: Startup Name, dtype: int64
```



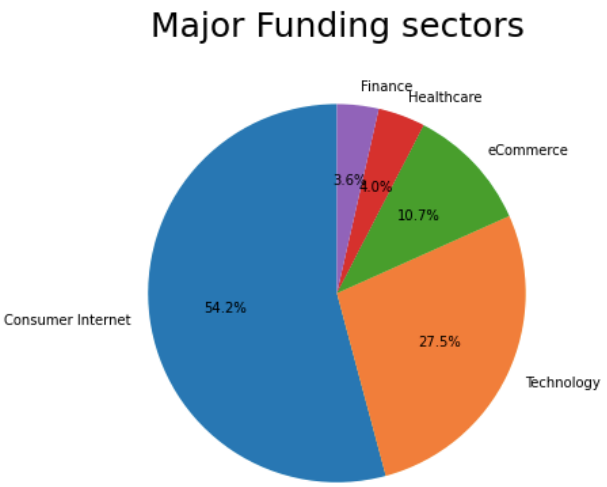
- Favourable Industries for Funding's

In [17]:
```
            #INDSUATRY VERTICLE
#Which industries are favored by investors for funding ?
#Which type of startups get fundings more easily?

industry = fd['Industry Vertical'].value_counts().head(10)
print(industry)
#here we displayed top 10 categories
```

```
Consumer Internet    941
Technology           478
eCommerce            186
Healthcare            70
Finance               62
ECommerce             61
Logistics             32
E-Commerce            29
Education             24
Food & Beverage       23
Name: Industry Vertical, dtype: int64
```

In [55]:
```
fundings = fd.groupby(["Industry Vertical"]).size()
plt.title('Major Funding sectors',fontsize=25,pad=80)
plt.pie(industry.values, labels = ['Consumer Internet','Technology','eCommerce','Healthc
plt.show()
```



- Cities having maximum start-ups

In [21]:
```
#which cities are having maximum startups.
fd.rename(columns={'City  Location':'City Location'},inplace=True)  #changing column nam
cities = fd['City Location'].value_counts().head()
fd['City Location'][fd['City Location'] == 'Bengaluru'] = 'Bangalore'
print(cities)
```

```
Bangalore    701
Mumbai       568
New Delhi    424
Gurgaon      291
Bengaluru    141
Name: City Location, dtype: int64
```

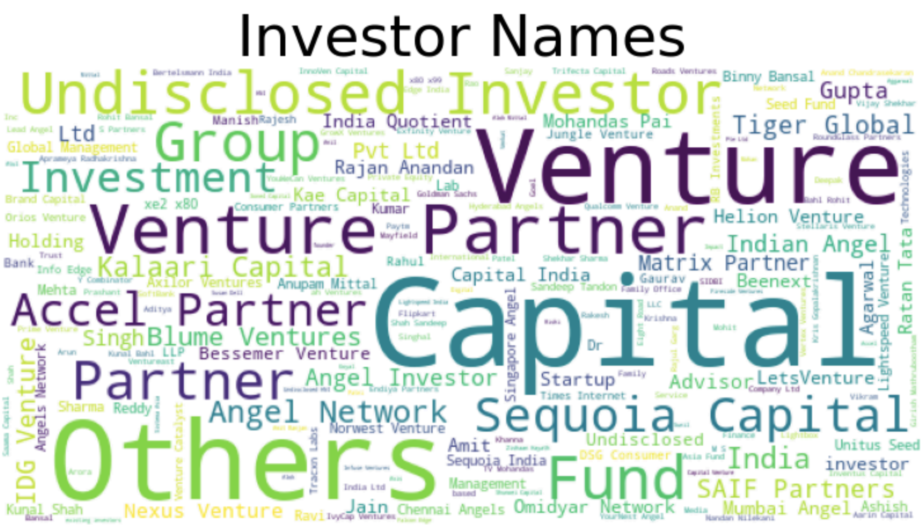
In [65]:
```
from wordcloud import WordCloud
names = fd["City Location"][~pd.isnull(fd["City Location"])]
wordcloud = WordCloud(background_color='white',max_font_size=80, width=600, height=300).
plt.figure(figsize=(13,8))
plt.imshow(wordcloud)
plt.title("Best Startup Locations", fontsize=35,pad=30,color='red')
plt.axis("off")
plt.show()
```

- Names of the most Important investors

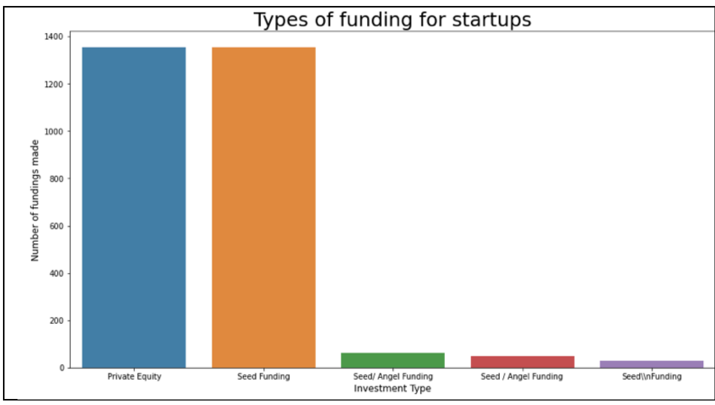
In [23]:
```
#Names of TOP 5 important investors
print("TOP FIVE INVESTORS:")
names = fd["Investors Name"][~pd.isnull(fd["Investors Name"])].head(5)
print(names)
```

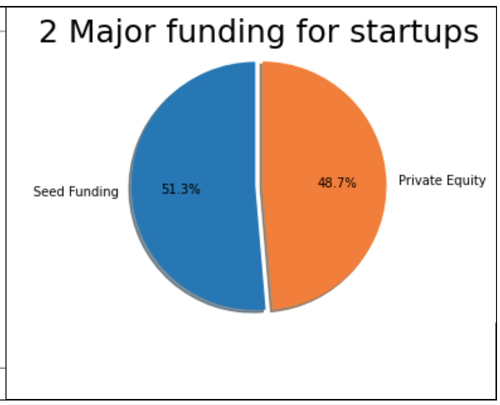
```
TOP FIVE INVESTORS:
0      Tiger Global Management
1     Susquehanna Growth Equity
2         Sequoia Capital India
3               Vinod Khatumal
4       Sprout Venture Partners
Name: Investors Name, dtype: object
```

In [68]:
```
names = fd["Investors Name"][~pd.isnull(fd["Investors Name"])]
wordcloud = WordCloud(background_color='white',max_font_size=80, width=600, height=300).
plt.figure(figsize=(13,8))
plt.imshow(wordcloud)
plt.title("Investor Names", fontsize=45)
plt.axis("off")
plt.show()
```



In [26]:
```
plt.figure(figsize=(15,8))
sns.barplot(investment.index, investment.values, alpha=0.9)
plt.xticks(rotation='horizontal')
plt.xlabel('Investment Type', fontsize=12)
plt.ylabel('Number of fundings made', fontsize=12)
plt.title("Types of funding for startups", fontsize=25)
plt.show()
```





- Confusion Matrix

In [136]:
```
#confusion matrix para
y=fd['raised']
X=fd.drop(['raised','Startup Name','Industry Vertical','SubVertical','City Location','In
print(X.head())
print(y.head())
type('raised')
```

```
   Sr No  Year  Amount in USD   Age  5years
0      1  2020    200000000.0  2020       0
1      2  2020      8048394.0  2020       0
2      3  2020     18358860.0  2020       0
3      4  2020      3000000.0  2020       0
4      5  2020      1800000.0  2020       0
0    2020
1    2020
2    2020
3    2020
4    2020
Name: raised, dtype: int64
```

In [137]:
```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=42)

scaler=StandardScaler()
scale=scaler.fit(X_train)
X_train=scale.transform(X_train)
X_test=scale.transform(X_test)
```

In [138]:
```
model=LogisticRegression()
model.fit(X_train,y_train)
pred=model.predict(X_test)
```

In [139]:
```
score=accuracy_score(y_test,pred)
score
```

Out[139]: 0.9605911330049262

In [140]:
```
confusion_matrix(y_test,pred)
```

Out[140]:
```
array([[179,   0,   0,   0,   0,   0],
       [  1, 193,   0,   0,   0,   0],
       [  0,   0, 140,   0,   1,   0],
       [  0,   0,   3,  66,   3,   0],
       [  0,   0,   0,  15,   7,   0],
       [  0,   0,   0,   0,   1,   0]])
```

In [141]:
```
matrix=classification_report(y_test,pred)
print('Classification report:\n',matrix)
```

```
Classification report:
               precision    recall  f1-score   support

        2015       0.99      1.00      1.00       179
        2016       1.00      0.99      1.00       194
        2017       0.98      0.99      0.99       141
        2018       0.81      0.92      0.86        72
        2019       0.58      0.32      0.41        22
        2020       0.00      0.00      0.00         1

    accuracy                           0.96       609
   macro avg       0.73      0.70      0.71       609
weighted avg       0.95      0.96      0.96       609
```

Github link to code implementation:

https://github.com/Ak27-18/Indian-Startup-Success-Analysis

10. Conclusion:

Finally, we arrive at the end of this analysis. Here we go through a step-by-step procedure to build a machine learning model to predict the high potential success of start-ups and to explain the model. We begin by understanding the goal, preparing the dataset, and creating the model. Then to understand the contribution of each input feature to the prediction results we used matplotlib and seaborn for visual representation and using confusion matrix calculated the accuracy of the model. This step-by-step method should give us an idea of creating and explaining our model results.