## **Start-up Analysis**

Contents

[Introduction 5](#_Toc38395875)

[Start-up analysis 5](#_Toc38395876)

[Motivation 5](#_Toc38395877)

[Problem Statement 5](#_Toc38395878)

[Literature Survey 6](#_Toc38395879)

[Solution Design 7](#_Toc38395880)

[Solution Approach 7](#_Toc38395881)

[Technology Stack 7](#_Toc38395882)

[Design Model 8](#_Toc38395883)

[Solution Implementation and Results 9](#_Toc38395885)

[Obtaining Data 9](#_Toc38395886)

[EDA 9](#_Toc38395887)

[Pre-Processing 12](#_Toc38395888)

[Machine Learning Algorithms Used 13](#_Toc38395889)

[Results 14](#_Toc38395890)

[Conclusion and Future Work 15](#_Toc38395891)

[Conclusion 15](#_Toc38395892)

[Future Work 15](#_Toc38395893)

[References 15](#_Toc38395894)

# Introduction

## Start-up analysis

A startup is a young company founded by one or more entrepreneurs to develop a unique product or service and bring it to market. By its nature, the typical startup tends to be a shoestring operation, with initial funding from the founders or their friends and families.

In the early stages, startup companies have little or no revenue coming in. They have an idea that they have to develop, test, and market. That takes considerable money, and startup owners have several potential sources to tap:

Traditional funding sources include small business loans from banks or credit unions, government-sponsored Small Business Administration loans from local banks, and grants made by non-profit organizations and state governments.

* So-called incubators, often associated with business schools and other non-profits, provide mentoring, office space, and seed funding to start Ups.
* Venture capitalists and angel investors actively seek out promising startups to bankroll in return for a stake in the company once it gets off the ground.

## Motivation

Coming up with brilliant startup ideas may feel tricky to aspiring entrepreneurs, especially when it may seem that everyone’s already swooped up every good idea for a business. Still, it’s entirely possible to become successful by improving on existing products or putting a unique spin on an old idea. The benefits of self-employment can make the effort of launching a startup worth it. In addition to the freedom that comes from being your own boss, starting a business brings more independence, greater job satisfaction, and potentially uncapped earning potential.

## Problem Statement

This project shows the insights of funding done by startups and how growth changed with several factors. The aim of project is to get a descriptive overview and a relationship pattern of funding and growth of newly launched startups. Another important point to understand how funding changes with time is an important aspect. Possible area of interests would be – (Funding ecosystem and time relation, cities as an important factor, which industries, important investors). Dataset we are using contains information of funding of startups from January 1980 to 2014.The amount invested is in USD. Aggregation of data w.r.t cities, investors, funding type etc. is required to get an optimized result. Here we done major pre-processing of data and overcome problem of missing data and uncertain distributions. Also, Visualizations are done to find the anomalies and mining patterns from data. It seems to be some cities showing some abnormal behaviour when it comes to funding.

# Literature Survey

As today's world is making big transformation by startups, on an average 90 out of 100 startups fail to attain the expected range of profits. . There can be several reasons like inefficient planning, inefficient way of using the funds, lack of good team to work, insufficient funds, etc. which leads to failure of startup. This work aims to create a machine learning model for predicting the range of funding that can be expected by the organization based on many important factors that play a major role in all stages of a startup. It is very important to increase the success rate of startups. It helps in predicting the future of startups by involving machine learning and increases the productivity and decrease the failure percentage. The main factor that influence the success of the startup is the idea which further gets implemented. A signiﬁcant number of startups fail because of lack of awareness about the progress of that particular domain. Every entrepreneur aims at development and success of his/her organization. They want their idea to be helpful and appreciated by their customers so that the organization will receive the enough profits to expand share in market and to make further modifications required. The main elements of any startup are: 1. Funding 2. Appointing best employees. 3. Reduce the employee attrition. This paper aims at building models for predicting the amount of funding the startup can receive based on the factors like product idea, domain, sub vertical to which it belongs to etc. It is desired to know the future of the project and it’s plausibility to give success far before the implementation of the idea. Much work has not been addressed to solve this problem. The machine learning algorithms will help to handle such cases more effectively. This work mainly aims in guiding to build the model which predicts the range of funding a startup can expect far before the implementation process.

The crunch base dataset us a raw startup data which includes different attributes of different types and it is consolidated data. Working with such datasets demand for rigorous data preprocessing. So, the previous work on this dataset focusses on data cleaning, data reduction, data transformation and normalization. The basic steps involved in data preprocessing are (a) Converting attributes that hold numeric values to symbols (nominal) for example funding type (seed funding, private equity etc.) (b) Constructing the attributes that represent time like seed funding date etc. The date attributes are replaced with the number of days from a fixed date so that the data remains integer for better comparison.

# Solution Design

## Solution Approach

Initially we started with the data collection that was done from Kaggle and various other sources. The raw data needed to be pre-processed to remove the NAN / Missing values. After data pre-processing, Exploratory Data Analysis (EDA) was performed to achieve insights and statistical measures from the data. With the help of graphical representations in the form of various charts, we found out the correlations between various attributes. By using various design model we can to know about various relation between funding rounds and different funding types. The algorithms that we decided to use for rating prediction are Linear Regression, Lifelines and time series for category classification, we have used Decision tree Logistic Regression algorithms.

## Technology Stack

Hardware

* RAM - 4 GB
* CPU – Intel i3/i5/i7

Software

* Python

Libraries

* + pandas
  + numpy
  + matplotlib
  + seaborn
  + sqarify
  + sklearn
  + sklearn.model\_selection
  + warnings
  + wordcloud
  + sklearn.linear\_model
  + sklearn.metrics
  + lifelines
  + LabelEncoder
  + Preprocessing
  + Utils
  + Graphviz

## Design Model

## 

As per our problem statement we need dataset of startup and investors companies so we found it online and collected. It contains records of last 50+ years and it is nicely detailed dataset but unclean. Applied several pre-processing steps: to make it consistent removed null values, redundant and unimportant data. Converting attributes to suitable data types and selecting important features. Our problem statement is more of analyzing patterns, trends and behavior of the funding ecosystem, we tried to it through visuals and graphs more.   
  
There is some important fields like market and category, investors, region, funding type level, total funding amount, and quarters and funding rounds which is selected for training and testing models. For model algorithms used are multivariate regression and decision tree. Multivariate algorithm works to predict funding amount needed and decision tree to categorize status of any startup company in future looking on provided data. As the data is more in string format and having many unique values, model is unable to get high accuracy score.

# Solution Implementation and Results

## Obtaining Data

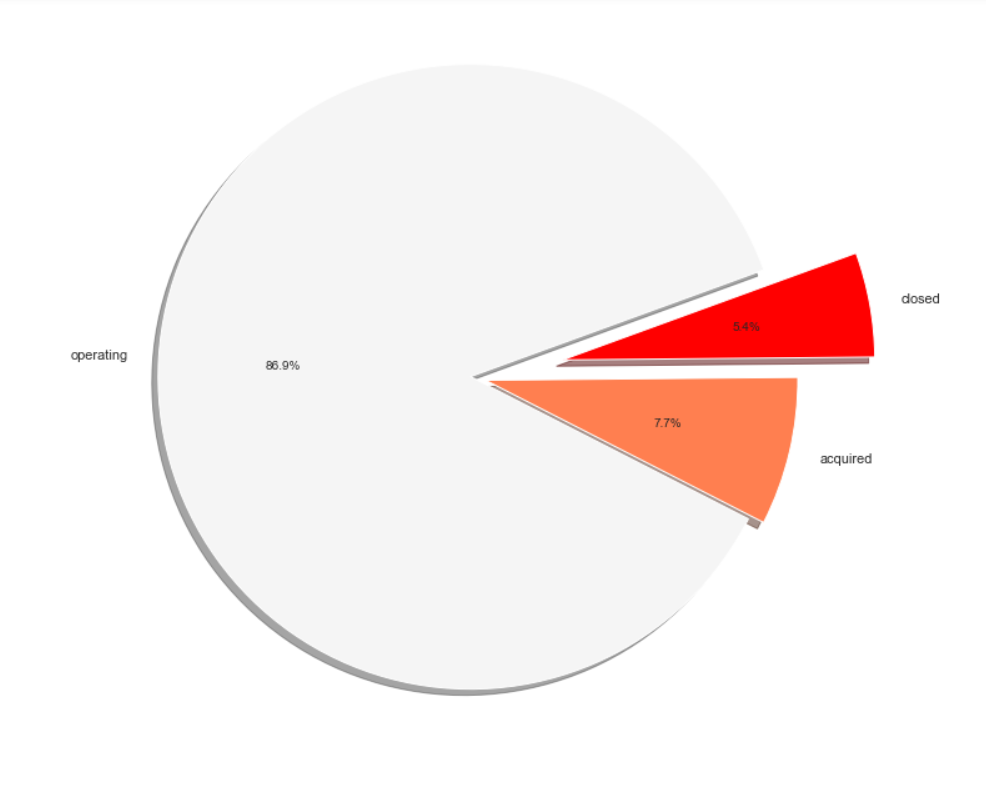
The crunchbase dataset us a raw startup data which includes different attributes of different types and it is consolidated data. Working with such datasets demand for rigorous data preprocessing.

There are various data sets which available on the various sources such as GitHub, kaggle UCI and Crunch base and Track.in

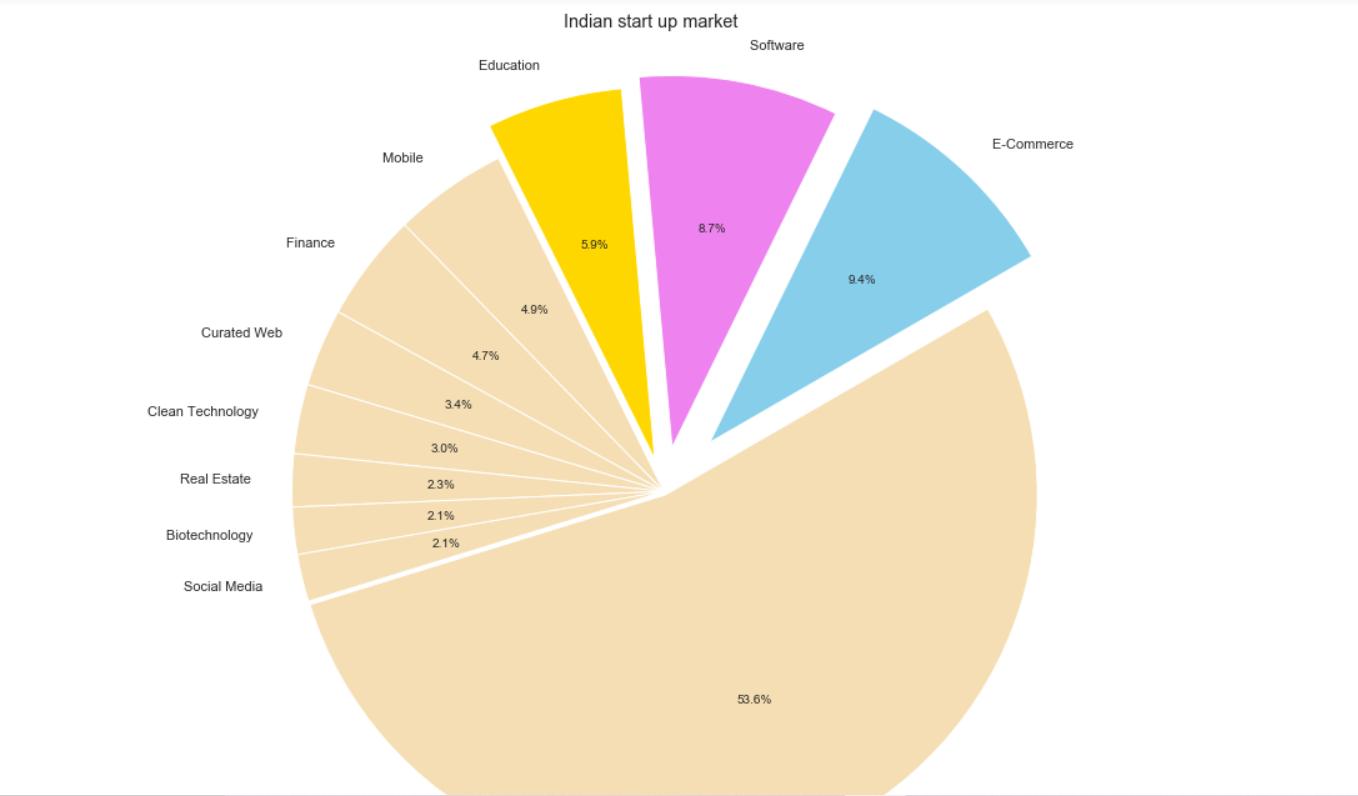
## EDA

**Exploratory Data Analysis** (**EDA**) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

EDA was done in the initial stage of our project, where we found interesting findings which shape the further work. Some of these are as follows

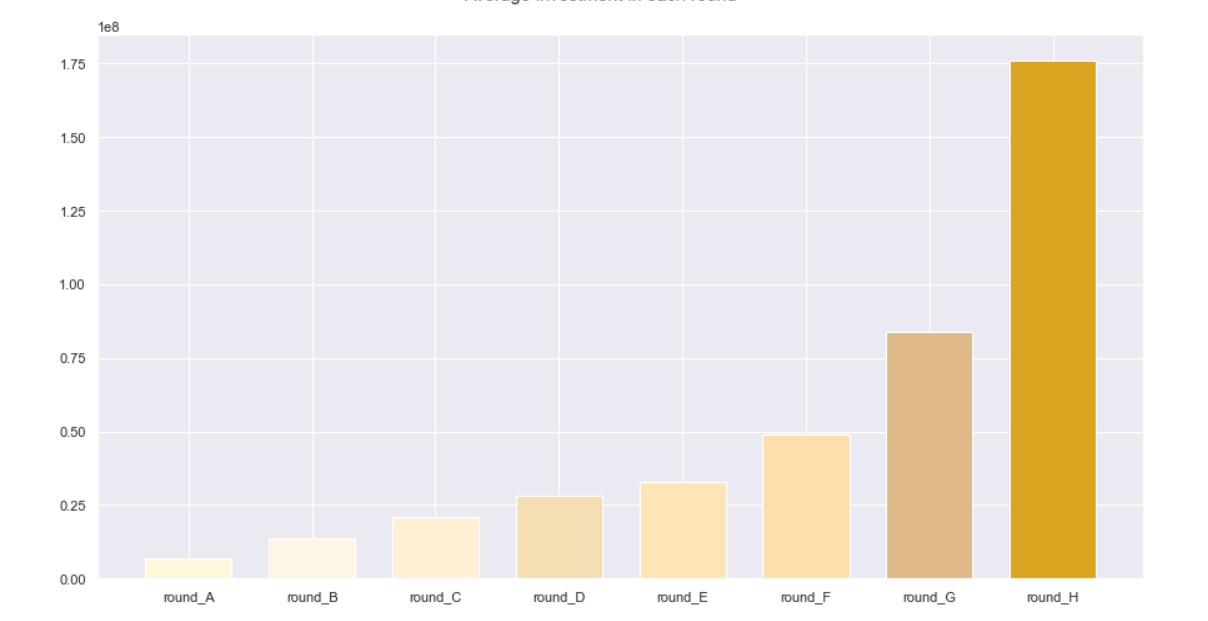


The above diagram shows the current status of start Ups around the world. The diagram is Pie chart which describe the current status for start Ups. The Start Ups which are operating are 86.9%, those which are acquired are 7.7% and are closed due to some reasons are only 5.4%.

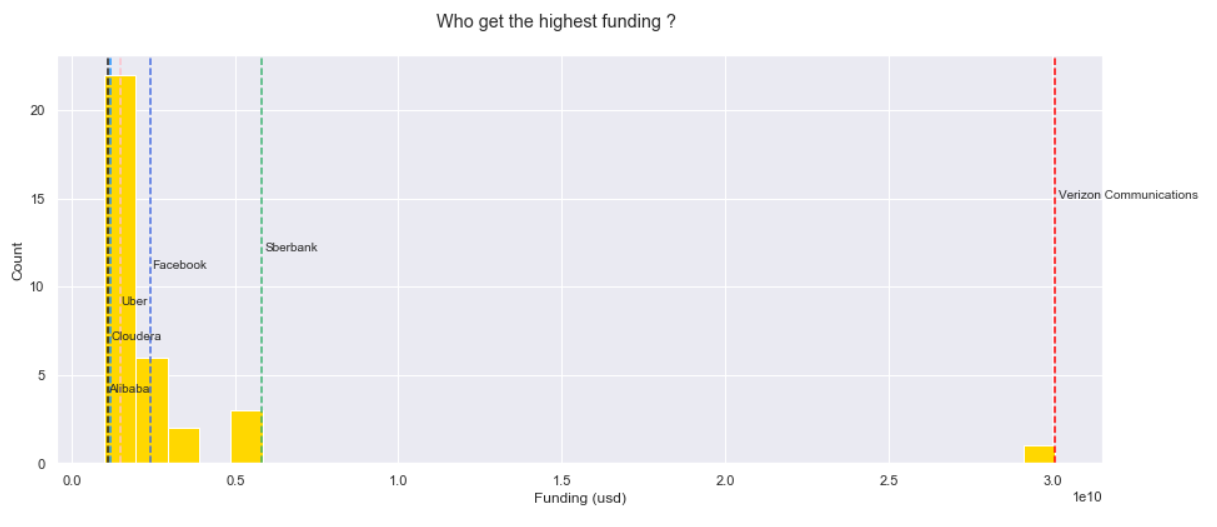


The above diagram is pie chart of the Indian Start Up market.

The diagram explains the majority share in Indian Start Ups is of e-Commerce and Software building which contribute to Indian Start Ups. The major 53.6% share is combined of all other small sectors which contribute from 0.1 to 1% share.

,

The above diagram explains the average investment in each round. The start Ups get the funding gradually which increases with each round. The mean of all funding rounds is calculated to build the bar graph.



The above diagram describe which company got the highest funding of all start Ups. The companies like Alibaba, Cloudera, and Uber received about $0.2 to $0.3 million while Verizon Communication received $3.0 million.

## Pre-Processing

Since the data was scrapped from the Crunch Base, there were null / missing values in the dataset. The raw data extracted needed to be pre-processed to turn it into some valuable information. To be able to perform EDA and run algorithms on our dataset some attributes were converted to integers. First we have selected data which is not null as there were null values towards the end (tail) of the data set so we have not considered it. Later the categorical attributes like Status were label encoded, to convert string into categorical values to perform further analysis. Also, we found that there are NAN values in various columns so have backfill method and other thigs to handle this problem and we have also used data scaling to scale and normalized the data.

## Machine Learning Algorithms Used

**[1]KNN:**

The k-nearest neighbours (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. To select the K that’s right for your data, we run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors we encounter while maintaining the algorithm’s ability to accurately make predictions when it’s given data it hasn’t seen before.

**[2]Time Series:**

A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future. Time series forecasting is the use of a model to predict future values based on previously observed values.

Time series is used to analyse a component or a value or a period of time. We have used this component to analysis funding of start up over a period of time. This is very use full in predicting trends over the year. In our project we have analysed what type have funding is most sought after is start-up want keep on operating or accrued.

**[3] Random Forest:**

The random forest is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses two key concepts that gives it the name *random*:

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

**[4]DBSCAN:**

Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning.

DBSCAN groups together points that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.

**[5] Decision trees:**

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Decision trees provide an effective method of Decision Making because they: Clearly lay out the problem so that all options can be challenged. Allow us to analyse fully the possible consequences of a decision. Provide a framework to quantify the values of outcomes and the probabilities of achieving them.

**[6] Logistic regression:**

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

## Results

* In linear regression were able to predict weather an amount in funding round can be calculated or not by venture funding.
* In Kaplan–Meier and coxwe were able to predict how start-up funding was required over a period of time to ensure that start-up does not closes down due to lack of funding.
* In time series we saw funding trends and how much was moved over a period of time.
* In classification algorithms we find out does a start-up can be in operation status if certain about of funding is given or not.

# Conclusion and Future Work

## Conclusion

It is seen that a very small number of start Ups succeed after the market launch of products or services and continues to develop and make profit. Start Ups, which are associated with high-tech projects, often seen losing its way from the founding stage of the start-up to achieving business success. Data on the number of start-up companies is based solely on the information gathered by searching the information received from Annual Funding Report. The dearth of earlier research against which the results of this study could be compared to is indeed a limiting factor. Despite these limiting elements, this analysis has shown start-up companies prefer primarily funding in different stages. The result has also shown that maximum funding was received in the Venture funding. Grant funding is becoming more and more prevalent.

With the visualizations, we came to know that various events do affect the funding. Also the cities and industry verticals play an important part in acquiring funding from investors, and foreign investors invest a lot into start Ups.

The result revealed that although funding of a start Up is primarily depending on the type of funding i.e.; seed funding and private equity, but also the other factor, which has been assimilated in this research as independent variable has a significant impact on the funding. In this study it also has been found that Developed cites are the best start-up ecosystem. This study also shows impact of market/Domain on start-up funding.

## Future Work

Providing In depth analysis of start-up funding and domain based funding on various types of funding and studding does city and domain affect start-up survival.

## References

List the references in IEEE format

1. Kumar, Mahesh. (2019). Growth Pattern and Trends in Startup Funding in INDIA. 8. 3721-3724. 10.35940/ijitee.L2654.1081219.
2. Hari, T. N. (2018). The Indian Startup Report 2018. Bengaluru: YourStory Media Pvt Ltd