Startups in India

Springboard Data Science Career track

Capstone 2

Patrick Long

3/29/2022

Overview

The purpose of this study is to gather past data on start-ups in India to better understand how, and where future companies will be funded.

Why is this important? From a macro level view, India has become the third-largest startup ecosystem in the world after the US and China. As of January 2022, India had 83 'unicorns' with a total valuation of $277.77 billion, according to a Forbes Survey. (In business, a Unicorn is a privately held startup company valued at over US$1billion). Most of these are in the service sector, which contribute to over 50% of India's GDP. The Survey noted that the government recognized over 14,000 new startups in 2021-22 against 733 in 2016-17. With this, the total number of recognized startups in the country surpassed 61,400. During 2021, 555 districts had at least one new startup against 121 districts in 2016-17, as per the Economic Survey.

During this time India saw a record 44 startups turning unicorns in 2021. Unicorns are companies with over $1 billion valuation. India also overtook the UK to have the third-highest number of unicorns after the US and China, which added 487 and 301 unicorns, respectively.

Although the pandemic of 2020 and 2021 stop a lot of funding, India was still able to supply its startups with funding. Private market investments, just as public market investments, thus, got a boost in India since 2020. According to VCCircle in December, reported that all unicorns of 2021, put together, had raised over $12.7 billion..

All of this makes it so important for investors, looking for an edge, should turn their attention to India. With startups in the US and China being very competitively fought over, startups in India might be easier to grab a hold of.

However, with all the new startups coming into place how can investors find the right one? I will lay out a case on where to find the best places to find startups in India.

I will use these factors to determine where and how start-ups are being funded:

-Amount funded

-Where the funding is coming from

-Which state the funding is going to

-Which industry is being funded

-When the start-ups are being funded

After having an initial look at the data, I formulated a hypothesis:

Most funding will go towards the most populated cities in India, Mumbai, and Delhi. There will be mostly service industry startups, and that they will be funded by private equity.

That will be because of government incentives, population, and a trend of startups to gather in cities for employee pools.

Data Wrangling

Datasets

I will leverage data from multiple sources. The data will come from Kaggle, RBI, and Statisticstimes.com.

https://www.kaggle.com/paree24/india-gdp-growth-world-bank-1961-to-2017

https://rbi.org.in/Scripts/PublicationsView.aspx?id=20006

https://www.statisticstimes.com/economy/comparing-indian-states-and-countries-by-gdp.php

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

The data comes in CSV form and will need to be cleaned and merged.

The first step of this project is the Data Wrangling section, found here:

https://github.com/ptlong11/capstone2/blob/main/Capstone%20Two%20-%20Data%20Wrangling%20(2).ipynb

The approach I took to data Wrangling was to first clean the data sources. That means deleting any arbitrary columns, finding the null values, and replacing or removing them.

After each set of data was clean, I merged them all on the 'State' column to end up with one data frame that had 'State' as the de facto index. This allowed me to work within that data frame comparing the funding amounts and types by various states. This approach helped me figure out my initial hypothesis.

The Kaggle data had GDP growth in India from 1962 to 2017. I used the 2017 year, to be as up to date as possible. The RBI data had GDP growth per state, which was broken up by year. I kept it to 2017 to be the same with the Kaggle data. There were no missing data in either set for 2017.

The next Kaggle data set is crucial to the project, as it contains the startup funding. The data has great integrity and only had a few missing values, which were removed during the data wrangling section.

During the data wrangling process, I merged the data sets on 'state.' Because 'state' was the most common column for each data set, it felt most appropriate. That way I would be able to compare the GDP per state and how much money was going into each state.

Some insights from the merged data set:

```
startup_state_gdp['State'].value_counts(normalize=True)*100
```

```
Delhi            94.949495
Chandigarh        2.222222
Goa               2.020202
Kerala            0.202020
Haryana           0.202020
Karnataka         0.202020
Uttar Pradesh     0.202020
Name: State, dtype: float64
```

This shows us the value count percentage of startups per state. You can see that Delhi has a vast majority of startups.

```
startup_state_gdp['Industry'].value_counts(normalize=True)*100
```

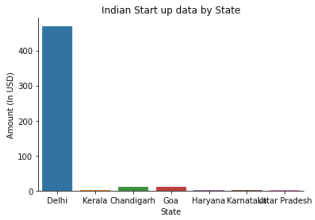
```
Consumer Internet                       31.313131
ECommerce                               13.535354
Technology                              12.929293
Healthcare                               3.030303
Food & Beverage                          1.414141
                                           ...
Online content platform for women        0.202020
Premium Beverages                        0.202020
Digital Healthcare                       0.202020
Logistics Service Provider Marketplace   0.202020
Transportation                           0.202020
Name: Industry, Length: 170, dtype: float64
```

Here, we can see that Consumer Internet has the highest percentage of startups at 31.31%. Next, is Ecommerce at 13.5%, and Technology at 12.9%. These two insights give a good indication that my hypothesis, that most startups will be in Delhi, be in the service industry, and will be funded by private equity, is correct.

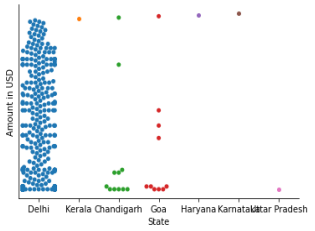
Exploratory Data Analysis

After Data wrangling, the EDA process takes place. This is where I will try to gather as many insights in the data as possible. That will mean that I will be attempting to find more insights on which state is receiving the most money. See where the money is coming from. See if there is a possible 'Funding Season' (best time for startups to raise funds during the year). I will also be running a Kmeans clustering analysis. That refers to a collection of data points aggregated together because of certain similarities.

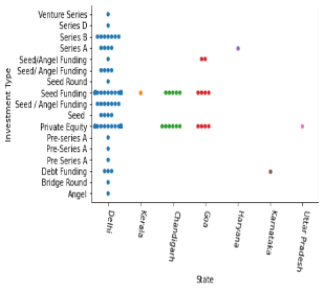
My first data analysis was to visualize the startups per state. As we know from the Data Wrangling portion most startups are in Delhi. However, let's look at what exactly that looks like:



This graph shows where each startup is located, by State. Though we can see again that Delhi has most of the startups, there still are many located in Chandigarh and Goa. This is great insight for those investors that really want to find a differentiating edge of their competition. If there is healthy startup economy in states other than Delhi, then, there may be great untapped investing opportunities that the competition will not find.

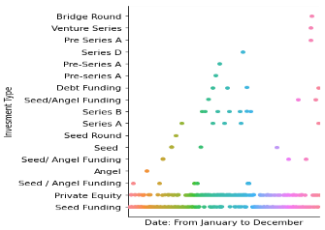


Now we turn our attention to the type of funding that is funding startups. The data of funding is broken down into types of funding. To name a few; "Venture series" "Series A, B, C, D" "Seed." These are examples of classic startup fundings. Private equity and Seed funding are the most popular type of funding across all of India.

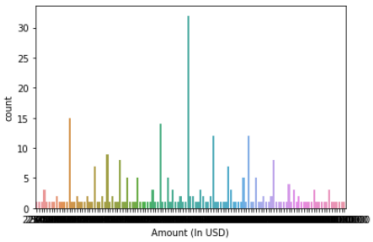


I wanted to see if there was a specific time of year that startups are looking for funding. Are startups getting more funding in the summer or winter? I wanted to make sure that any investor who wanted to use this data to find startups in India was as well-equipped as possible. This data below gives us an indication that there are not any specific time periods. It is much more so reliant on the type of funding.

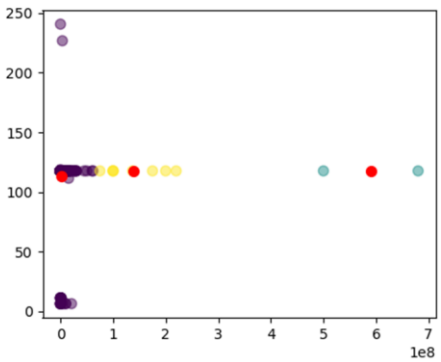
Startups in India



I wanted to find if there were any 'Unicorns' of my data set. So I took a look at all of the funding of each startup I am analyzing. We can see that there are 10 odd startups that are well above the average. Though none are in the billions of fudning. That doesn't mean they aren't unicorns. As we know, the valuation of a company determines that. As far we know, these 10 could be Unicorns!



I ran a Kmeans clustering analysis between GDP (which is broken up by different states, so we will be able to tell which state is which) and Amount in USD (which is the different amounts of funding per each startup).



The state with the GDP right above 100 is Delhi. That is the state with 117 Billion in GDP. The state at just above 0 on the chart is Goa with 6 Billion in GDP. This shows us what I predicited and what we've sen previously. Most startups being funded are in Delhi, with a few outliers that show up as significant.

Startups in India

Pre-Processing

Purpose: Now that I have obtained, cleaned, and wrangled my dataset into a form that is ready for analysis, it is time to perform exploratory data analysis (EDA). I will Keep in mind that the goal of the EDA work is to get familiar with the features in your dataset, investigate the relationships between features, and understand the core characteristics of my dataset. Be creative and think about interesting figures and plots you can create to help deepen my understanding of the data.

Goal: Create a cleaned development dataset you can use to complete the modeling step of your project.

With this project I Created a dummy indicator features for categorical variables

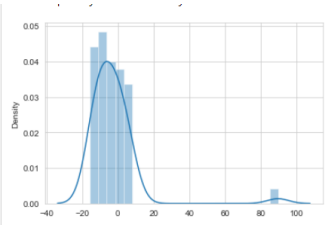
Steps I will be taking in the process I will be:

- Creating dummy or indicator features for categorical variables
- Standardizing the magnitude of numeric features using a scaler

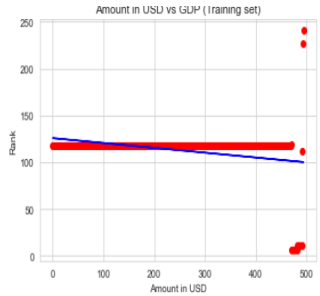
  I will be answering the questions 'Does my data set have any categorical data, such as Gender or day of the week? Do my features have data values that range from 0 - 100 or 0-1 or both and more?'

During this process I will construct a regression analysis and run it through multiple training sets. A random Forest is a meta estimator that fits several classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

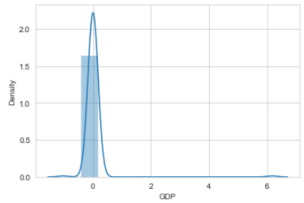
First, I split the data of the GPD and Amount in USD into independent and Dependent variables. I then divided the complete dataset into training and testing data. There I found the Density of the data:



Here is the training set, with the predictive model in blue:



The random Forest Regressor state was set at 1. We can see that our previous observations were in fact correct. The vast majority of funding of all kinds are going into Delhi.



Modeling

Purpose: The goal of the modeling step is to develop a final model that effectively predicts the stated goal of figuring out how much money, where its coming from, and where to invest in Indian start-ups.

I modeled that Amount in USD and the GDP Rank of each state. Meaning I wanted to dive deep into which state was receiving the most money. That way investors who want that extra edge in finding startups to fund can either dive deeper into those states or avoid them.

First step was to find an accuracy score using a Logistic Regression analysis. The higher the score means the model is performing better.

```
The accuracy score achieved using Logistic Regression is: 8.62 %
```

Second, I wanted to find the accuracy score using a Linear Support Vector machine. The score came out to the exact same as the previous test.

```
The accuracy score achieved using Linear Support Vector Machines is: 8.62 %
```

Third, I ran a Naive Bayes accuracy test:

```
The accuracy score achieved using Naive Bayes is: 5.17 %
```

Fourth, I ran a K Nearest Neighbors test:

```
The accuracy score achieved using KNN is: 3.45 %
```

Finally, I ran a Decision tree accuracy report:

```
The accuracy score achieved using Decision Tree is: 5.17 %
```

So, what does this mean? The higher the accuracy of this means the model is performing better. This score is quite low comparively, which speaks mostly to the data not being very varied.

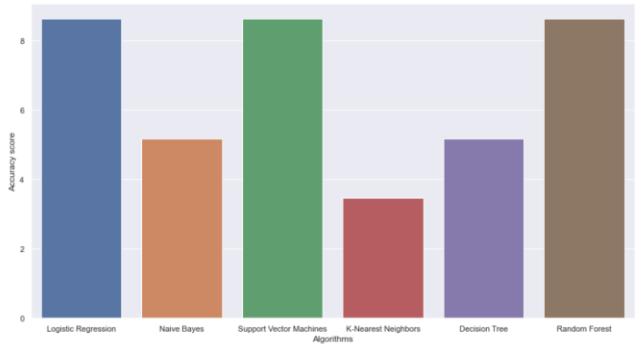
The SVM came out to 8.62%, which again is not high.

The Naive Bayes test is a kind of classifier which uses the Bayes Therom. It predicts membership probabilites for each class such as the probility that given record or data points belongs to a particular class. The score is 5.17%, which again is very low. If the data set were larger we would see a larger number here.

The K Nearst Neighbor test is used for pattern recognition tasks. It is a supervised learning algorithm that can be used for regression as well as classification problems.

The Decision Tree is a non=parametric learning method used for classification and regression. This accuracy report suggests much of what the other modeling reports have. The data isnt that numerous.

Here are all six of the modeling I ran grouped. Ultimately not a lot can be gained from these test, mostly because the data wasn't quite that numerous, and Delhis had nearly all of the startups.



Startups in India

Findings and Recommendations

Ultimately, trying to find and predict the best state in India to invest in is hard. Because so much of the start-up industry is in one state, Delhi, that is the de facto place to invest. My initial hypothesis was that Delhi would have the most startups with the service industry being the most numerous industry and private equity being the most popular funding method.
Those claims were quick to come to fruition, however the project was more revealing than that. What I was able to find is that yes, most startups are in Delhi there is a lot more to it. There are enough startups in other states to make it significant.
The investor that is looking for an extra edge would be prudent to start in a sector like Technology, in a state like Goa. Because there are still enough startups there to make researching and funding worthwhile, it will not be flooded with other investors like Delhi.

Why didn't the modeling work? This happens from time to time in Data Science. There are a few reasons that it could be. For this case it is data quality. That is not necessarily saying the data I had was poor. There just was not that much data. This can lead to those exceptionally low accuracy scores that were seen on each model. It could also be an example of overfitting. The data did not have much variance, so, the models were not able to 'learn' anything.

I would recommend any that is doing a project like this one. I.e., looking for startups in India to start with this project as the framework. Now that we know where most startups are coming form, how they are funded, and in what industry, I would search elsewhere. Look for government incentives to start and fund startups. Or look for other ways that might indicate where new startups might emerge. I recommend that because as we can see there is still room for growth outside of Delhi, but to know when and where that growth will come is the next challenge.