Startup Analysis

Final Report

Vibha Satyanarayana[*], Sruthi V[†]
Department of Computer Science, PES University
Bangalore, India
Email: [*]vib.satya@gmail.com, [†]vsruthi98@gmail.com

***Abstract*—**

This project was done to analyze startup dynamics and understand the various factors regarding startup funding. We obtained data from kaggle[11] and also scraped data from other sources. A series of cleansing processes were done on both datasets. We then used different visualization techniques to view and understand the data before any models could be built on them. In this process, we made a few interesting observations and tried to reason out the same. Then, classifiers were constructed on this data. Different ones were experimented with to both produce good results and accommodate the features of this particular dataset, like the large number of categorical features present.

I. Introduction

Startup entrepreneurship is crucial because it brings innovations, new jobs and competitive dynamics into the business environment. Most of present day freshers and employees who are laid off turn towards startups. But they have no insights on current market trend, need for this startup in their city,viability of the market for the service they're providing and the major investors around them. Apart from knowing their skills and aptitudes, knowledge about their competitors in the same domain can help them determine what their ideal working environment is moving forward.

In the recent years, massive growth in startup economy of India has attracted investors all around the world. But statistics[9] show that 9 out of 10 startups tend to fail.Top reasons [3]. for these numbers is lack of market need, no investor interest and being out competed. This paves way for the requirement of the entrepreneur to know what are the factors that will affect his or her startup. Startups fail when they are not solving a market problem,and are instead tackling an interesting problem,which goes on to show the importance of market analysis location wise.

Our idea was set up on this framework. We intended to produce solutions for these problems by analyzing data on startups in the year 2015 to 2017. Our methods will help look at the bigger picture, see where every city stands in terms of startup economy, who are the key investors in play and it might also help potential founders to make decisions wisely. Data regarding startups in India is extremely sparse or mismanaged, and hence we also extracted data from websites[4][1].

II. Literature Survey

Many approaches were seen in this arena out of which these were of interest to us as they aligned to our problem perspective.One of the works[8] discusses predicting the outcome of a startup based on factors such as seed funding amount and seed funding time. The authors have included a ranked list of positive and negative severity factors as an influence on the outcome. A total of 9 models were built experimenting on various classifiers and layering an additional seed funding type for every higher model.The next approach[6] seen focused on crowdfunding. Some of the notable attributes that were scraped or calculated are the number of Facebook friends, Twitter followers and sentiments. Yet again various classifiers were experimented on, inclusive the additional component, AdaBoost[13].

Another paper [12] focuses on estimating the relative importance of a variety of approaches and variables in explaining pre-startup success. The authors derived possible success and risk factors from dimensions such as characteristics of the individual(s) who start the venture, the organization which they create, the environment surrounding the new venture, etc. They showed how full time startups are more successful than part time ones. In this work [7], the authors focused on the cognitive abilities of the individuals receiving funding. A standard multiple regression was performed between the amount of funding secured as the dependent variable and predevelopment meticulousness, social influence, risk aversion, and learning orientation.

Work[10] done on the same dataset as ours includes various visualizations for data interpretation.

III. Problem Statement

We intended to address the issue of startup funding and suggesting investors for startups, based on their industry vertical. Budding entrepreneurs will benefit by knowledge about their peers in the same sub vertical as theirs, the type of investment that is popular for their sector and the range of investment that this particular domain receives. We propose to look at the bigger picture by visualizations, see where each city stands in terms of startup economy, who are the key investors in play and help potential founders to make decisions.

A. The dataset

The original dataset[11] that this project is built on is from the **website kaggle**. This dataset contained information on the industry vertical and subvertical the startup belonged to, date of investment. They key details were the amount of funding the startup received, from which investor and the type of funding.

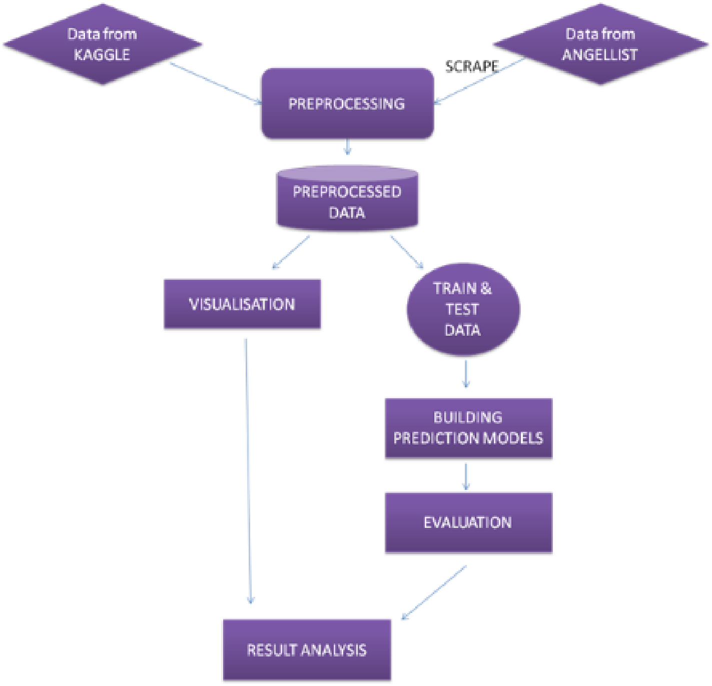
Apart from the existing data, we have managed to **scrape data** regarding the startup founders from other websites. This data gave an outlook on founder credentials, such as educational qualification, university information, skill set and other factors.

The **preprocessing** required for both the above mentioned datasets was quite extensive, as the data was sparse and the text containing columns were not of proper format. This, along with the fact that the data is from the 2015 to 2017(peak years for startups), were constraints imposed on our approach.

IV. Proposed system

A block diagram (Fig 1)summarizing our approach

Fig. 1. Flowchart depicting the steps of the proposed system



V. Components of the system

Here, we have discussed the components of our system in detail, including the various experiments done on them and the results obtained under each.

A. Preprocessing

The initial dataset had to be cleaned column wise, by first removing missing value rows from our key column, Amount of funding. The next step was to format the date column and extract correct locations from the City Location information. Redundant values were cleansed and the data was made ready to use.

Since the education of startup founders is was in focus, the initial step in this dataset was to extract all the universities from the given list for each row. We also labeled each candidate as a founder or employee by analyzing the designation field.

The data in all attributes was categorical. In order to use this to predict funding, they had to be fused into smaller categories and encoded. The encoding was done by assigning ranks to each category . The same was done to the amount attribute as predicting amount would be very specific, difficult to do and highly inaccurate. All our models predict the tier of the company (a category with a specific range of funding values).

B. Visualization and Results

Various techniques and approaches were taken to vizualise the data and understand the setup. Plotting the data helped us know the arena we were working in and also choose the questions that had to be answered about the data wisely.

- **Startup count over the these three years** (2015-2017 August) was observed. The increase that was seen in the second half of 2015 gradually dropped with the advent of 2016, which might be an effect of the results of US Presidential elections.
- We then wanted to take a close look at every year and see **the industries that appeared in funding discussions the most**. Though the prominent industry in all three years was Technology and Mobile the point of interest was that startups of the food category had gone down since 2015, and was seen to lose its second place to E commerce over the years. And reasons[2] were that market was already saturated with a number of these.
- Analysing the **startup economy of India** by viewing the map showing different cities, we can easily spot that the most number of startups is produced by Bengaluru and Mumbai.
- To find top Investors in major Industry Verticals and suggest the same to founders, we picked the most frequently occurring industries and visualized the **investors that totally funded the maximum amount**. This also helped us see the **range of investments** each industry was receiving. For example, the maximum funds received by education startups is 20 million, while the same in Technology is 300 million.
- Visualizing the average and maximum startup funding per month, the immediate glaring observation were four peaks, two in 2015 and 2 in 2017. These belong to the startups Flikart and Paytm. While in 2015 they collected in the ranges of 600 Million, they became what is known as the **”Unicorn startups”** in 2017, with about 1.4 billion each. Unicorns refer to companies whose present valuation has surpassed the mark of $1 billion.
- We then moved on to the founder credentials data. On a closer look, we discovered that the startups from this dataset had very less to do with the ones from our original dataset. Still, with the spirit of analysis, we continued to inspect and visualize this dataset too. Under this, we first tried to understand the **roles that universities played with producing startup entrepreneurs**. It could be seen that the the maximum number of startups originated from various Indian Institute of Technology, with Delhi and Kharagpur in the lead. Further scrutiny showcased that the number of startup founders and startup employees from all these universities were almost equal.
- The founders’ skills and market specialization was used to create a **wordcloud** to show the **current status** of the country in terms of business and technology . Some examples include- python, java, product development, consumer internet etc. The investors wordcloud depicts the top investors with maximum contribution. Some of the top investors are Tiger Global, Sequoia Capital, Soft Bank Group, Microsoft and eBay.

C. Predictive model

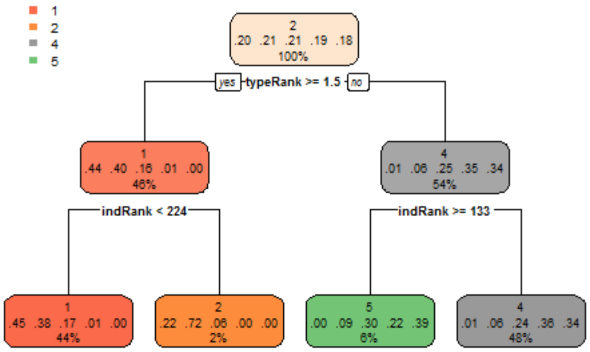
Test and train data

The dataset was shuffled in order to get a random sample. This was divided into 70% training data and 30% testing data.

The dataset has four major attributes that could affect the amount attribute. These are - Industry Vertical, City, Funding Type and Month (the other attributes were not considered due to a large number of *na* values). Scatter plots of these attributes and amount were plotted in order to determine their correlation.

- **Linear Regression** : The amount column was divided into **five categories (tiers)**, each one having a range of fund values. Linear regression was used to **predict which tier a company would belong** to given its industry vertical, its location, funding type and the month it started. The **AIC value** for this model

Fig. 3. The below figure shows the probabilities of every tier as outcome.



was found to be **2177**. As we can see from the above graphs, month is not correlated to amount. Hence month was dropped to improve the model. This gave an AIC value of **2175**. Predictions using this model gave an **accuracy of 33%** and a root mean square error value of 1.18. This low value of accuracy could be due to the fact that the dataset has only categorical data and linear regression works best with numerical data. Also, the model could be improved by using better string to numeric encoding techniques. Linear regression to predict the amount directly was not given importance as it gave very high root mean square error .

- **Multinomial Regression** : MLR extends the binary logistic model to a model with numerous categories (in dependent variable). However it assumes that there is no order in the category to which the outcome belongs.
  The **AIC value** of the model was **2027**. Again, month was dropped and the new model was compared with the previous one. The AIC value was **2020**. The confusion matrix was built using the test data and the **accuracy** was found to be **40%**. **Root mean squared error was 0.97**. This model is better than the linear regression model as it is built for categorical data. The performance could be improved by better encoding.
- **Ordinal Regression** : Ordinal Regression is also an extension of binomial logistic regression. Ordinal regression is used to predict the dependent variable with 'ordered' multiple categories and independent variables. This model is more realistic as it considers the natural order in the categories. Hence it is **theoretically, the right model for this dataset**. The **AIC value** of this model was **2013** with an **accuracy of 41.5%**. The root mean squared error was 0.96. As expected, the accuracy is slightly more in this model compared to the previous one.
- **K means clustering** : The dataset was divided into 5 clusters, one for every tier. The

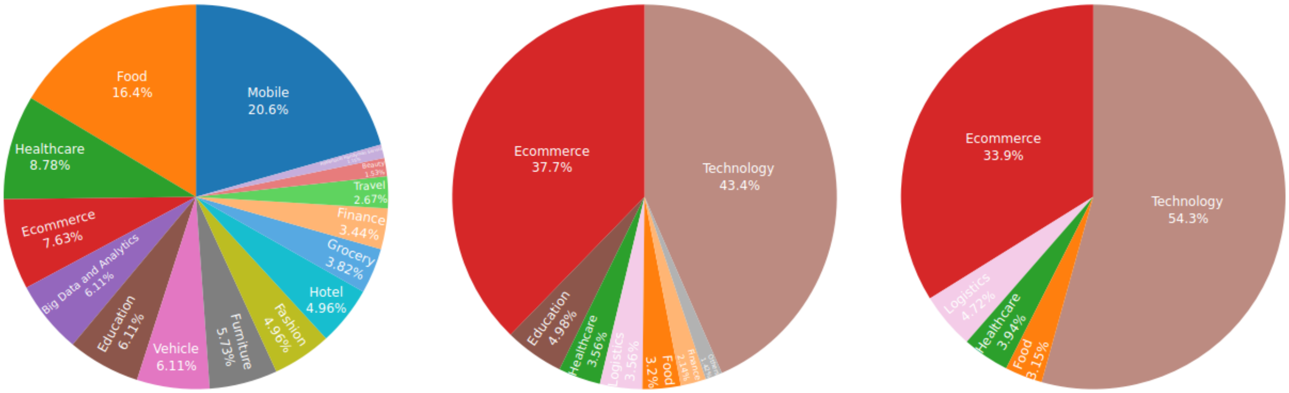


Fig. 2. Industry vertical split up in the years**(1)2015 (2)2016 (3)2017**

Euclidean distance was used as the distance measure from the centroid. The root mean squared error got from this model was 2.1 - a high value probably due to the fact that k means is sensitive to outliers[5] and this dataset had a large number of them (apart from other reasons mentioned above).

- **Decision trees** : Decision trees are an effective method for decision making as they provide a framework to quantify the values of outcomes and the probabilities of achieving them. This **model is easier to interpret**. Hence it is the best model for entrepreneurs (who are non-statisticians) to understand and make informed decisions. The accuracy got from this model is 40.67% with an root mean squared value of 0.96.

VI. Conclusion

We began the project by extracting data from various resources. Extensive preprocessing made us realize the crucial role of valid information in data analysis. In order to construct predictive models visualization was the first step. These graphs made us question some glaring observations and it was necessary to reason them out.

After experimenting with various classifiers, we came to a conclusion that the ordinal regression model was the best for this dataset. The results clearly show that the outcome would have been better with a larger and cleaner dataset and if more external factors were taken into consideration.

Startups are an integral part of the country's economy. Thus, analyzing startup dynamics will not only be useful to individuals who founded them but also to the nation as a whole.

VII. Contribution of each member

Vibha Satyanarayana
(USN : 01FB15ECS346)

- Scraping data regarding founders from various websites
- Cleaning scraped data
- Building various models with different classifiers and experimenting with the same to produce good results

Sruthi V
(USN : 01FB15ECS311)

- Cleaning original dataset to make data ready to use
- Cleaning scraped data
- Plotting different graphs and analyzing the same by reasoning out various observations made.

References

[1] Angel.co.

[2] Decoding: Why food tech startups are not successful in india?

[3] The top 20 reasons startups fail, cbinsights.

[4] Trak.in.

[5] A unified approach to clustering and outlier detection.

[6] Michael D. Greenberg, Bryan Pardo, Karthic Hariharan, and Elizabeth Gerber. Crowdfunding support tools : Predicting success and failure. 2013.

[7] Peter A. Koen, Gideon D. Markman, Robert A. Baron, and Richard Reilly. Cognitive mechanisms: Which ones allow corporate entrepreneurs to obtain startup funding. 2005.

[8] Amar Krishna, Ankit Agrawal, and Alok Choudhary. Predicting the outcome of startups: Less failure, more success. 2016.

[9] Neil Patel. 90% of startups fail: Here's what you need to know about the 10%.

[10] Jaghadish Rajagopalan. Simple exploration notebook-indian startups.

[11] Sudalai Rajkumar. Indian startup funding.

[12] Marco van Gelderen, Roy Thurik, and Niels Bosma. Success and risk factors in the pre-startup phase. 2005.

[13] Wikipedia. Adaboost.