PREDICTION OF STARTUP FUNDING

by Nama Naga Hitesh

Submission date: 07-Dec-2021 11:10AM (UTC+0530)
Submission ID: 1723111942
File name: Batman-PlagCheck.pdf (216.95K)
Word count: 2222
Character count: 10909

PREDICTION OF STARTUP FUNDING

Team name : Batman

Team members :

J Jeshwanth Reddy - PES2UG19CS154

V Manas Advaith - PES2UG19CS908

Nama Naga Hitesh - PES2UG19CS244

Suhas R - PES2UG19CS413

I. Introduction

The dataset contains information about various startup companies in various cities of India. It also has information about the investment type, industry vertical and the amount funded for them. Our task is to predict the funding amount for the testing data. Prediction of this funding amount varies corresponding to the investment type and the city in which the startup is located. Funding amount prediction can help a lot of upcoming startups to plan their proceedings accordingly. Without knowing the funding amount, there can be many issues with it which lead to the failure of startup. Main aim of our work is to build a machine learning model that can predict the funding amount for startups. This plays a major role in increasing the success rate of startups by predicting their future. It decreases the failure percentage of startups by involving machine learning.

This paper aims at making use of machine learning algorithms to predict the funding a startup can receive and this helps them to plan accordingly since it is given before the process is implemented. The prediction depends on various factors such as city, investment type etc. We used the concepts of find S algorithm, random forest classifier and linear regression to predict the funding amount. A summary of previous work is presented below.

II. PREVIOUS WORK

We have used machine learning techniques which are mostly supervised and unsupervised. Ensemble models like random forest and unsupervised model like find-S. We also used linear regression with data encoding to predict funding values by matching one axis. All the coding has been done in python and the data set is free of outliers and vague, unavailable and nan data.

a) Find S algorithm

Find S algorithm comes under machine learning, it is a concept to search and match categorical data. The attributes of a column are searched based on the categories present in the dataset to look for accurate and similar ones in the dataset. Find S stands for “find similar”. we took top 4 similar ones based the on the distance and averaged them all to encounter outliers.

b) Random Forest

Random forest is basically an ensemble model which is used for regression and classification problems. It is a supervised algorithm. It is based on the ensemble model of bagging and it uses the randomness of the features, it works by building many decision trees and then combing them based on e.g Voting. It is not prone to overfitting which gives us good confidence when we are building a model prone to this. Its default hyper parameters are tuned very well and give appreciable results. It requires more computation power as the number of nodes increases.

c) Linear Regression

The linear regression model is built by forming a linear relationship between independent variables to find the dependent variable. It draws a line that best fits all the points in the graph and that line is used to predict dependent variable.

LIMITATIONS

The main limitation for our dataset is that there are many categorical variables. So we are supposed to give more preference to classifiers rather than regressors which gave us less accuracy than expected. We tried the best possible ways to increase the accuracy while using classifiers.

ASSUMPTIONS

There are many categorical variables in the dataset. For linear regression model, since the regressor takes only numerical values, we encoded few categorical data columns as numbers. Finally, we train the model using these encoded columns and predict the funding.

III. PROPOSED SOLUTION

a) *Pre-Processing*

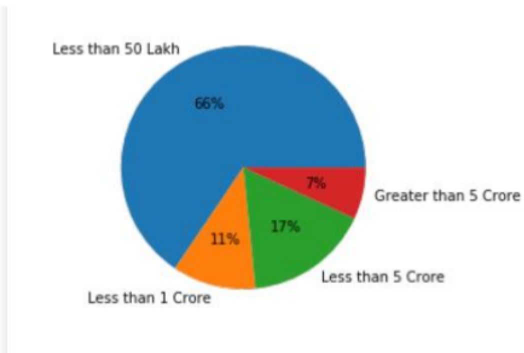
Data Pre-processing is a very important aspect of analyzing .We may never analyze a complete dataset and may miss out on the noise or the unclean data which will cause variation and bias in the model. Data pre processing may include Standardization, Normalization or dealing with missing values by imputing them or replacing them or by taking the mean in order to preserve data so as to get a considerable accuracy..

For our dataset, we used Data Cleaning and Data Reduction techniques. Rows having at least one null value are deleted since they are categorical variables and cannot be replaced by mean of that column. Two columns which are irrelevant and are not useful for predicting the funding amount are deleted.

b) *Data visualization*

Data visualization is a representation of data using various graphs like charts, bar graph, histogram etc. It is useful in understanding the data visually and is easy to check for trends and outliers in any.

First Visualization :- Start-up Funding Trends

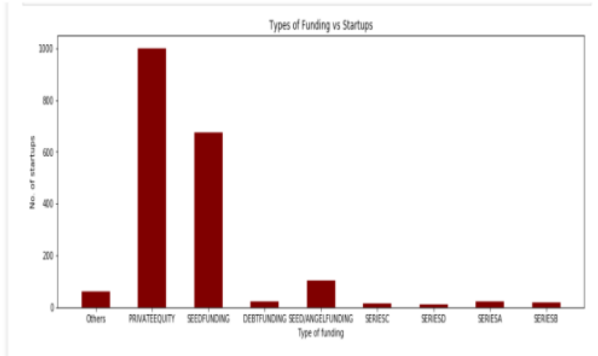


1) It is seen that, in the chosen dataset around 66% of the start-ups are allotted with less than 50 lakh rupees.

2) We can also see that only around 7% of the start-ups are funded with more than 5 crores rupees.

3) Also start-ups with less than 1 crore and less than 5 crore funding are 11% and 17% respectively. 4) From this chart, it can be observed that most of the start-ups are funding with less that 50 lakh rupees.

Second Visualization :- Types of Funding v/s No of Start-ups



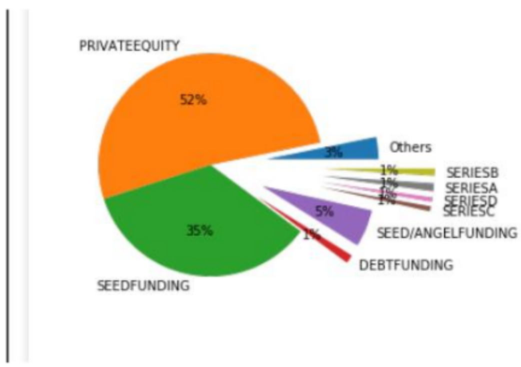
1)From the above graph we can see that majority of the start-up funding is from Private Equity.

2) Also a fair amount of start-ups are funded by Seed Funding type of funding.

3) Most of the other start-ups are funding in small portions by Debt Funding, Angel Funding, Series A,B,C,D types of funding.

4) So here most of funding is from either Private Funding or Seed Funding with the former being a majority.

Third Visualization :- Types of Funding v/s No of Start-up

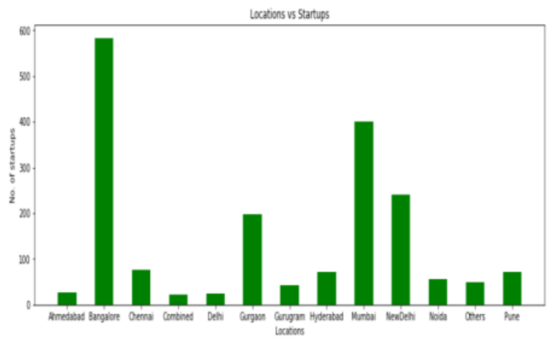


1) From the above graph we can see that around 52% of the start-up funding is from Private Equity. 2) Also a fair amount(35%) of start-ups are funded by Seed Funding type of funding.

3) Most of the other start-ups are funding in small portions by Debt Funding, Angel Funding, Series A,B,C,D types of funding.

4) So here most of funding is from either Private Funding or Seed funding with the former being a majority.

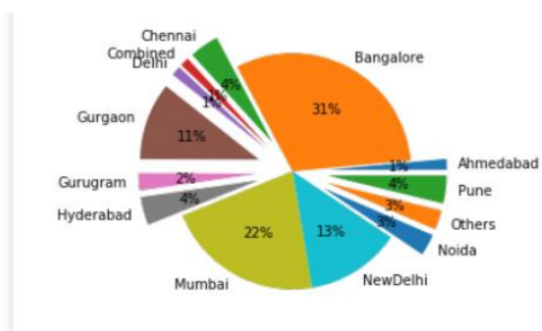
Fourth Visualization :- Types of Funding v/s Start-up Location



1) From the above graph, we see that most of the Start-up are emerging from lot of developing city like Bangalore, Mumbai, New Delhi, Gurgaon .

2) Also we can infer a lot of start-ups are founded from these cities.

Fifth Visualization :- Types of Funding v/s Start-up Location



From the pie-chart we can see that

1)31% start-ups from Bangalore

2)22% start-ups from Mumbai

3)13% start-ups from New Delhi

4)11% start-ups from Gurgaon

c) *Building the model*

1) *Random Forest*

Random forest is basically an ensemble model which is used for regression and classification problems. It is a supervised algorithm. It is based on the ensemble model of bagging and it uses the randomness of the features, it works by building many decision trees and then combing them based on e.g Voting. It is not prone to overfitting which gives us good confidence when we are building a model prone to this. Its default hyper parameters are tuned very well and give appreciable results. It requires more computation power as the number of nodes increases. Here we are performing Random Forest to find the amount for our given test data. We achieved this by label encoding the categorical features and then fitting the model with the relevant features and our goal is to predict the amount for the given features.

2) *Find S algorithm*

Find S algorithm comes under machine learning, it is a concept to search and match categorical data. The attributes of a column are searched based on the categories present in the dataset to look for accurate and similar ones in the dataset. Find S stands for “find similar”. we took top 4 similar ones based the on the distance and averaged them all to encounter outliers.

Here we are using the find S Algorithm to predict the values. Here we are storing the training and testing data in a list and then implementing the Algorithm. In this model we are considering the features city, industry vertical and amount from the training data to predict the amount for the given test data containing the features city, industry vertical .

3) *Linear Regression model*

The linear regression model is built by forming a linear relationship between independent variables to find the dependent variable. It draws a line that best fits all the points in the graph and that line is used to predict dependent variable.

This model uses linear regression by taking in all the relevant features i.e. 'Industry Vertical', 'City Location' and 'Investment Type' doing prediction on them by label encoding then so a to fit the model. Our Aim is to predict the amount for a test data containing these feature values.

d)*Model Evaluation*

We use three models that are Find S, Random Forest and Linear regression. We see that our Find S model has the best predictions as shown by the Mean Absolute percentage error. The random Forest gives us a lot of error and shows us that it is not fit for handling a data like this and moreover it requires a lot of computation power and time and is quite inefficient in this scenario. The Linear regression model has considerable amount of error but not as much as the random forest error and it is also moreover requires lesser computation power and can be said to be better than the random forest model. Label encoding has enabled us to fully use the capabilities of our dataset to the fullest to extract and predict to a very close value. The Find S model predicts very efficiently and effectively.

IV. RESULTS

The results for the model building for training are as follows:-

1) *Find S algorithm*

```
error= 9.988932803911418 %
accuracy= 90.01106719608859 %
```

Out[22]:

	Actual	Predicted
0	2000000	6000000.000
1	200000	4550000.000
2	1000000	3840000.000
3	120000	3220000.000
4	165000	2783571.429
5	150000000	21185625.000
6	2200000	19076111.111
7	10000000	14384307.692
8	25000000	15142571.429
9	400000000	40799733.333
10	2000000	36129176.471
11	50000000	34966368.421
12	900000	33263050.000
13	25000000	32869571.429
14	10000000	31830045.455
15	1000000	29231708.333
16	1250000	28112440.000
17	5000000	27223500.000
18	35000000	26886107.143
19	10000000	26593700.000

This model is giving high accuracy since it finds a specific hypothesis that fits all the positive examples. We have to note that the algorithm considers only those positive training example. The find S algorithm starts with the most specific hypothesis and generalizes this hypothesis each time it fails to classify an observed positive training data thus giving high accuracy. Here we take the mean of the most matching data with respect to our training dataset and then predict the value.

2) *Random Forest*

```
Root Mean Squared Error: 116045113.8479745
Mean Absolute Error: 30116290.374677002
Mean Squared Error: 1.346646844798936e+16
```

Random Forest is mostly used for classification tasks. It can also be used for regression tasks. Random Forest can't extrapolate. It can only make a prediction that is an average of previously observed labels. In other words, in a regression problem, the range of predictions a Random Forest can make is bound by the highest and lowest labels in the training data. This behaviour becomes problematic in situations where the training and prediction inputs differ in their range and/or distributions. This is covariate shift and it is difficult for most models to handle but especially for Random Forest, because it can't extrapolate. Therefore, we got high RMSE value and this model is not fit for predicting funding.

3) *Linear Regression model*

```
Root Mean Squared Error: 45237358.660685755
Mean Absolute Error: 23517965.13657666
Mean Squared Error: 2046418618595520.5
```

Here, Linear Regression model is giving a high RMSE value. Usually, LR model gives low RMSE value but that was not the case here. This is because, the variables in this dataset are categorical and we encoded them as numbers. The prediction would be more accurate only when there are numerical variables.

V. CONCLUSIONS

From the results , it is evident that Find S model is the best for predicting startup funding. We saw that Random Forest and Linear Regression model are giving high error value whereas Find S is giving an accuracy of 90.01% . Therefore, the best fit model is Find S.

The reason why the other 2 models are having considerable error is because our data has no correlation and this is evident by the PCA analysis. Since Random Forest and Linear Regression are regression models they work well with correlated data and since our data is uncorrelated we are getting considerable amount of errors .

a) *Contributions* **:-**

Jeshwanth Reddy – Random Forests and PCA

V Manas Advaith – Find S algorithm and managed team

Nama Naga Hitesh – Linear Regression model and data cleaning

Suhas -- PCA and EDA visualization

b) *Working Experience* :-

It was a very good learning experience for us to work with various machine learning models and learn insights from them. It was fun to evaluate each model and then try methods to improve the accuracy and decrease the error rate.

c) *References* :-

i) A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi. Exploration of data science techniques used to predict the strength of steel and Integrating Materials and Manufacturing Innovation, 3(8):1 – 19, 2014.

ii) Breiman. L. for Random forests. Mach. Learn., 45(1):5 – 32, Oct. 2001.

iii) kaggle.com for startup funding data.

iv) Pesu academy pdfs for information about machine learning models.

PREDICTION OF STARTUP FUNDING

ORIGINALITY REPORT

8%	7%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	towardsdatascience.com Internet Source	4%
2	www.ieee.org Internet Source	2%
3	Eliganti Ramalakshmi, Sindhuja Reddy Kamidi. "Predictions for Startups", International Journal of Engineering & Technology, 2018 Publication	1%
4	Submitted to Gitam University Student Paper	1%
5	export.arxiv.org Internet Source	1%

Exclude quotes On

Exclude bibliography On

Exclude matches < 5 words