RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

USING DATA SCIENCE CONCEPTS AND PRACTICES TO UNDERSTAND AND CLASSIFY THE STARTUP FUNDING ECOSYSTEM IN INDIA

Shiva Basava P[*1], S. Ush[a2], Vishwanath S Rao[3] & Sumukha K N Adiga[4]
[*1,3&4]UG Scholar, CSE, RRCE, Bangalore, India
[2]Prof. HOD, CSE, RRCE, Bangalore, India

ABSTRACT

Data science has become an important field of study and implementation across the world, to help customers Monetize the data that they possess effectively. Before establishing the startup we have to examine whether the company developed in that ecosystem will leads to any profit to us. We have to analyze all various factors before we invest on it. Such that we will perform the set of actions to predict whether the particular region or startup will help us to growth in ecosystem. We will do all the things by using data science concepts. Before the company loss we will ensure that no loss occurs before investing the huge amount on it.

Keywords: *-Data science, A/B testing, CHAID, Ecosystem, Startup.*

I. INTRODUCTION

Data science uses the theoretical, mathematical, algorithms and other practical methods to study and evaluate data. The key objective is to extract required or valuable information that may be used for multiple purposes, such as decision making, product development, trend analysis and forecasting. By making use of the data science we can predict or we can get the outcomes for a given problems. In simpler words data science can also be defined has an area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data.

A startup is a young company that is just beginning to develop. Startups are usually small and initially financed and operated by a handful of founders or one individual. These companies offer a product or service that is not currently being offered elsewhere in the market, or that the founders believe is being offered in an inferior manner.

An Ecosystem, in the context of a startup, refers to all the elements outside of the entrepreneur/s which have an impact on the success the startup enterprise.

In the context of a startup enterprise the following are some of the well-known elements of a Startup Ecosystem – apart from the core people forming part of the startup itself that is,

a. Universities and Research Organizations – providing access to - qualified manpower, facilities, research output / researchers, advisors and mentors
b. Big Companies – providing access to – experienced manpower, customers, advisors and mentors.
c. Funding Agencies – providing access to funds at various stages of evolution of the startup – angel investors who invest at the very initial phases of the startup, venture capital investors and private equity firms who provide funding during later phases of the startup leading to its acquisition by a larger enterprise or its going in for an Initial Public Offering.
d. Support Organizations – Like Incubators, Accelerators, Co-working spaces etc.
e. Service Organizations – Like Legal, Administrative,
f. Event Organizers – which conduct periodic events to bring together the various constituents of the ecosystem. Among all these, the most important element of the ecosystem are the Funding Agencies and the role that they fulfill in ensuring that the Startup survives and thrives – in most cases it is the difference between success and failure of the Startup. The Funding Agencies are of different types and focuses. The categorization is given below –

a. Angel Investors – these are people who will invest money right at the initial phase of the startup journey (because of this reason, this funding is also called 'seed funding'). Also more importantly these investors end up being individuals who have a good understanding of the domain that the startup is focusing on and can provide a lot of inputs into the product idea, the business plan and risks ensuring that the startup focuses on the right things and have a greater chance of success.

Venture Capital Firms – these are firms which invest money after the angel investment is over and the Startup has already has traction and is making some revenue. Thus these firms provide investments which will help the Startup to Scale and Grow – Geographically, Product wise or any other scaling model. Most of the times these Venture Capital firms specialize in specific areas of business (Like IoT, AR, FinTech etc.) and have built a lot of expertise within their company to provide support to the Startup company – which is in addition to the money that they invest. They also will help the Startup Company to either sell itself to a larger company or help it to go for a public issue.

II. DATA SCIENCE METHODOLOGY

Data Science may consist of five major steps. They are:

- Problem to Approach
  - Data Requirements and Collections
  - Data Understanding and Preparation
  - Modeling and Evaluation
  - Deployment and Feedback

a) Problem to Approach- This is the first step in which it consists of Business Understanding and Analytic Approach. In Business Understanding we have to understand the needs of business or problem of the situation which has to be addressed. In this stage we have to define the problem objectives that has to be achieved and solution to the problem from business side. This is the toughest stage we have to analyze the problem. In Analytic approach after the completion of business problem the data scientist analyzes and solve the problem by using various approaches. The approaches can be statistical methods or machine learning. Depending upon the desired outcome the techniques are used.

b) Data Requirements and Collections-Once the approach or model is build we have to determinewhich data is required for that particular model. After the required data is identified, the data can be either stored on structured or unstructured format. We have to collect the required data. Database management system is used for structured data. Unstructured data can be stored in Hadoop tool.

c) Data Understanding and Preparation- Data understanding involves the retrieving the data. Data preparation is preparing the data from the set of data. When the data is collected it can contain duplicates or it may also have unwanted data in it. The data has to be cleaned & delivered for further process. Data Cleaning assures that data must be valid, complete, consistent, uniform and accurate. This stage is the time consuming has the data to be refined.

d) Modeling and Evaluation- Modeling is a process in which data has to fit in the models. Evaluation is the process in which model is created is checked whether it has meet the business problem. This is the stage at which we can analyze the solution to given problem.

e) Deployment and Feedback- After the model is well developed it has to be deployed in the environment so that we can calculate the performance of it. Deployment helps to provide a set of new data and wok on it. By collecting the suitable results from model, the organization tries to upgrade the model again and again till it reaches the desired outcome of the business. The feedback helps us to provide the exactness and helpfulness of model.

III. TOOLS

There are different tools used for completion of data science some of them are-

1) R-R is a tool for statistics and data modeling. The features of R language are:

R is an interpreted language. So we can access it through command line interpreter.
R provides a large, logical, consistent and coordinated collection of tools for data analysis and also it is an open source software.

It has effective data handling and storage facilities.

2) Python- Python is used in data science because it has many number of built in libraries such has NumPy, SciPy, Matplotlib and Pandas. The features of python are:

It can be used for procedural and object-oriented programming.
Python is of freely available in nature, hence it is easy to run the program in any platform.

It is also known as Interpreted language

It provides interfaces to all major commercial databases.
3) SAS (Statistical Analysis System)-Statistical Analysis System is a leader in analysis business. Through analysis of innovative idea, it provides desired

requirements to business and management of data and business services. The features of SAS are:

This allows a user to make linear models, compute variance on different measurements as well as multivariate analysis on different parameters.

SAS software concentrates more on graphical analysis and requires less programming or coding knowledge.

4) SQL (Structured Query Language)-SQL Query language, it is more used in RDBMS. And still where so much traditional enterprise data resides. The features of SQL are:

- Generates scripts for data movement.
- It mainly makes use of DDL (Data Definition Language) and DML (Data Manipulation Language).
- By using commit and rollback functions we can back-up the data and recover the data.

5) Hadoop- It is an open-source framework that allows to store and process large unstructured and structured data in a distributed environment across group of computers by making use of simple programming models. The features of Hadoop are:

- Data Replication is unique features of HDFS. Replication solves the problem of data loss in an unfavorable condition like hardware failure.
- Hadoop is highly scalable in the way new hardware can be easily added to the nodes.

IV. ALGORITHM USED

1. CHAID Algorithm

The technique was created by Gordon V. Kass in year 1980. According to this, We discovery the relationship between variables. The variables may be one or more. It is based upon adjusted significance testing and used for detection of interaction between variables. This decision tree technique was developed in South Africa and is used for prediction and classification. Expansion of CHAID – "Chi- squared Automatic Interaction Detection". It is the higher version of AID (Automatic Interaction Detection) and THAID (Theta Automatic Interaction Detection) procedure. It graphically displays multivariate relationship and its tree structure output is easy to get answers. It makes use of multiway splits by defaults, for larger sample sizes of customers group it work effectively and the reliability analysis is more than the smaller sample sizes. This algorithm is good in uncovering structure, within the conditional and unconditional relationships among response and predictor variables. This algorithm chooses an

independent variable (predictor variable) that has the strongest relations with the dependent variable. Categories of each independent variable are merged if they are not significantly different with respect to the dependent variable.

The algorithm proceeds as follows:

Preparing predictors
The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

Merging categories
The second step is to cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable; for classification problems (where the dependent variable is categorical as well), it will compute a Chi-square test (Pearson Chi-square); for regression problems, F tests. If the respective test for a given pair of predictor categories is not statistically significant as defined by an alpha - to-merge value, then it will merge the respective predictor categories and repeat this step. If the statistical significance for the respective pair of predictor categories is significant, then it will compute a Bonferroni adjusted p-value for the set of categories for the respective predictor.

Selecting the split variable
The next step is to choose the split the predictor variable with the smallest adjusted p-value, i.e., the predictor variable that will yield the most significant split; if the smallest (Bonferroni) adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node. From above general steps we are deriving the following 5 Steps. Dividing the cases that reach a certain node,

Step 1- Cross tabulate the response variable with each of the explanatory variables. As shown in the Table 1.

Table 1-Tabulation of Experimental Variable

	NT=0	NT ≥ 1
Bad		
Poor		
Good		
V.Good		

\

Step 2 - This is applied to each table with more than two,

Table 2- Sub-table of Variables which is more than two.

	< 700	700-750	700-750	750 >
Bad				
Poor				
Good				
V.Good				
	X_1^2		X_2^2	

for each allowable sub-table. As shown in the Table

2. Comparing ($X1^2$) for the smallest X2 value. If it is not sufficiently great, later assemble the column groups. We have to repeat this Step 2 until we get till we get modest method of table.
Step 3 - Let us groups collective at step 2 to be splitted apart. For each compound category consisting of at least 3 of the original categories. We are following these sub-steps,

1. We have to find the greatest important binary distributed variable.

ISSN 2348 – 8034
Impact Factor- 5.070

2. If X2 is important, implement the split and return to Step 2.
3.Otherwise we are retaining the complex groups for this variable, and we are going to the following variable.
Step 4 - We have now finished the best combining of groups for each descriptive variable. Later we should discovery the greatest important of these optimally joined explanatory variables. As show in the Table 3,

Table 3- Categorizing Explanatory Variable

	C1+C2	C3	C4+C5+C6
Bad			
Poor			
Good			
V.Good			

Step 5 – We make use of the "greatest important" variable in step 4 to splitting the node with respect to merged groups for that mutable. As shown in Table 4

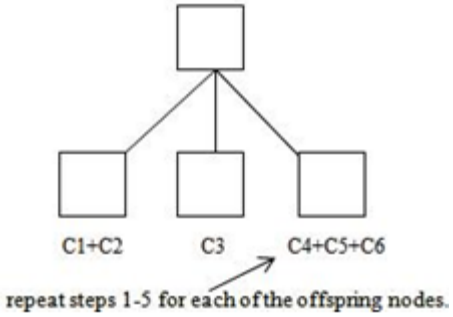

Table 4- Merging of Variables

At last we have to terminate, if no variable is important in step 4. The number of cases getting a node is below a specified boundary.

2. A/B Testing
A/B testing is used to compare two or more variables. It will help to govern the performance better among two variables. We can regulate the experiments with two trials of variables. It is a method of statistical hypothesis testing used in the area of statistics. A/B testing is a mode to match two methods of a single variable logically by testing a topic's reply to one variable against another variable, and finding which of the two variables are more effective in operation.

a. A/B testing Tools
A numeral tools are existing for A/B testing, with different focuses, price points and feature sets. Here are some:

Optimizely, SiteSpect, AB Tasty, Google Analytics. Optimizely: It is the experimental tool in which we make use of the features like "What You See Is What You Get" (WYSIWYG). It is also portable for experimentation to perform operation in various platforms.

SiteSpect: It mainly offers server-side solution and we can access the data from both the sender and receiver side. The customer experience and high performance for speed allowing for understanding view.

AB Tasty: It is an easy and good to start with this tool for beginners. AB Tasty software provides a good graphical user interface.

Google Analytics: It is the most frequently used analytical tool for analytics. It analysis the data from a mutual place. So that then we can share our incites to entire organization.

b. A/B Testing Process

The following is an A/B testing framework you can use to start running tests as shown in the below Figure 1,

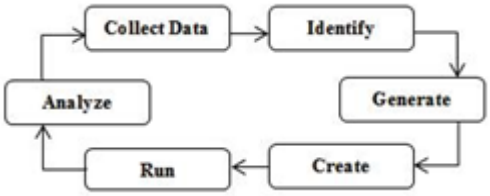

Figure.1-A/B Testing Process Diagram

1. Collect Data- The required data are selected from the datasets which are taken as a sample and given as the input for the A/B testing process i.e., Collect Data. This is the first step of the A/B Testing process. We are collecting the data so that it should provide the best result for predicting about the startup company establishment as well as for investing on that particular startup company.

2. Identify Goals- The main aim or goal of the paper is to predict whether the particular startup company can survive in particular region. And also the main objective is to predict the success or failure of that startup company, so the investors can invest the money on the company and they can also be in profit.

3. Generate Hypothesis- After the goals of the experiment or the project are being stated its time to make assumptions (or Hypothesis) on the datasets and to use the datasets on the operations of the process of A/B Testing Process. Based on the datasets, the hypotheses are being made and we are trying to make the better version of the current version. If we find the better version than the current version we are using the updated version of the output of generated hypothesis for the next A/B Testing process steps.
4. Create Variations- Now it is time for making variations in this step of A/B Testing Process. By making use of A/B testing software (like Optimize), we make the required changes that is of profitable. In this step we also perform some question and answer type of operations to check for our variations or modification which we had done by the end of the hypothesis.
5. Run Experiment- In this step we have to run the experiment and we have to observe the control of the process. We will compare, measure the dataset during running process
6. Analyze Results- Once the process is complete we have to analyze or compare the result with other data. If there is a variation in the data then it has to be iterated again and again and improve the result. If the data has the negative value or no result then the data is constant by using the learning we can generate new assumptions that we can run. The output of the process can be used for further tests and it can be iterate to increase the profit.

V. OVERALL PROCESS

In the below diagram (Figure 2) we can depeict the overall process or working of the system after it is built.

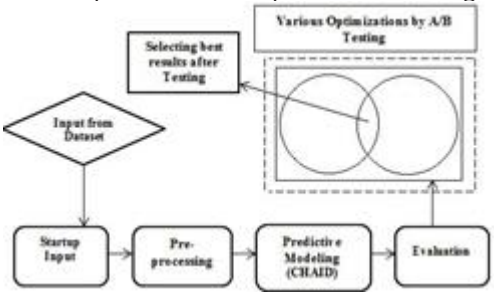

Figure 2- Overall Process

From the above (Figure2), It represents the overall process that is carried out to predict the success and failure of the startup company in the particular region by the help of the input dataset. After getting the required data for the process we are arranging them in particular order in order to give that output as input to the pre-processing stage. In pre-processing stage of the system it will process the data as that is required for predictive modeling i.e., It will arrange the data according to p-value. So that we can construct the model as per the CHAID algorithm. Now its time for the algorithm to sort the data according to the p-value (as discussed in the above Algorithm Section). Later the modeling of the dataset will be done by the end of the process of this CHAID algorithm. The dataset which is set from pre-processing will be now reconstructed by the CHAID algorithm in a tree fashion. It's time for evaluating the process that is, are we getting the right hypothetical result by plotting the graph. And we have to compare the results. For doing so, we are approaching to the final stage of the system process i.e., carried out by optimization by A/B Testing phase of overall process. At this stage we will be having all the results, but we are cross- checking the results by undertaking them to certain testing methods. Here the two samples are being tested of the same evaluated results (i.e., previousstage). The reason behind to consider two sample of the evaluated results is that, we will consider the common or best part of the both tested results. So that we will get the optimized result and we can ignore the other variables result that is making the result less\ effective.

VI. CONCLUSION

By using Data Science concept we are trying to make the prediction such that a startup company can survive in particular region so that a company and private investing company can make a profit. We have used one of the Decision tree algorithms that is CHAID algorithm which performs multi-level splits for computational purpose of classification trees that constructed the dataset in the manner that we are required it to be in the form for evaluation.

This method of predicting establishment of startup company can only be applied at the early stage of the startup. So that it can be a better method for referring and we can get the track of the startup company. Getting track is that, we can predict its market standings if it has followed this method for its establishment. It will also provide certain precautionary measure that the company should follow in order to not to commit the mistake or recurring or repeated mistake. We have also used A/B Testing method to test our results that we get after the CHAID algorithm process. This testing process is an added advantage to the company that it is mandatory that it should undergo in order to maintain its effectiveness in its establishment and also it's standing in market.

REFERENCES

1. *Amar Krishna, Ankit Agrawal, Alok Choudhary, "Predicting the Outcome of Startups: Less Failure, More Success." International Conference on Data Mining Workshops IEEE 16th 2016*
2. *Kohavi, Ron; Longbotham, Roger (2017) "Online Controlled Experiments and A/B Tests".*
3. *M. Ramaswami, R. Bhaskaran "A CHAID Based Performance Prediction Model in Educational Data Mining" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010*
4. *In Sammut, Claude; Webb, Geoff. Encyclopedia of Machine Learning and Data Mining (PDF). Springer.*
5. *Kohavi, Ron; Thomke, Stefan (September 2017). "The Surprising Power of Online Experiments". Harvard Business Review: 74–82.*
6. *a b c "The ABCs of A/B Testing - Pardot". Pardot. Retrieved 2016-02-21.*