Project State Contribution in Kickstarter Crowdfunding: A Look In to Data Based on Business Perspective

Saad Hasan

Department of Systems and Computer Engineering,

Carleton University, K1S 5B6, ON, Canada

***Abstract*—This research project is based on the idea of crowdfunding as this idea is old but as the time of innovation and information technology dragged this idea to business. So, for the crowdfunding. Three things required one is the idea, second is the platform to collect funds while third is number of backers to support the project based on their interest. In this project, we are using machine learning supervised algorithm for the classification of project include successful, failures, live and canceled with the help of modified artificial neural network to learn the pattern of the crowdfunding data. We also found best category to start the project to get the project successful through visualization techniques. In the last part, we used gradient boosting machines algorithm to predict the multi classification in down sampled data with same proportion of state as of the whole dataset. Kickstarter is an American platform responsible for the crowdfunding idea and gained billions of dollars in pledged amount form different people and organizations. Finding the pattern in the crowdfunding will help the business to improve not for the large platform but also for the small ideas to change from an idea to reality.**

***Index Terms*—Supervised machine learning, Classification, Kickstarter, Crowdfunding**

I. INTRODUCTION[1]

In the latest trend, Crowdfunding developed a thrilling alternate to the old source of business and project financing. It refers to the efforts by entrepreneurial individuals and groups – cultural, social, and for-profit – to fund their ventures with the help of internet, we can gather contributions which could be slight or gigantic but if we ignore standard financial mediators then we have less contributions [1].

While it is distinct in nature, it could have uncertainty like traditional funding sources linked with it; founders and funders are both connected with the success of the project. Here “success” for the founders is to achieve their funding goal and start on the project. Hence, it is important to understand what data signals the funders about the project’s worth and persuade them to invest. This project investigates data of about 300,000 projects on Crowdfunding website Kickstarter. It seeks to find patterns based on product type (e.g. music, food, technology, games etc.), fund goal, number of backers, project deadline etc. which could help us out to learn which projects are going to be profitable.

Now the question arises what is crowdfunding. Crowdfunding is the way of getting money from people to fund the project. There are many ways to get your project funded based on the interest of the people. There are three main people in the crowdfunding. First the person or persons who have the idea or a proposal to start the project, then the other most important people are the people who sponsored the project based on interest and the third most important individual is the platform to support the idea to raise the money or getting funded by the people to support the idea or project.

There are two main types of crowdfunding, first one is reward sourcing while the second one is equity crowdfunding. Reward based is that when someone gives an idea or business idea to launch without his or her investment so, in this type the project is totally dependent on the people funding. In the equity crowdsourcing, the project, the examples include scientific creation, research, or creation of a software. The second type is Equity crowdsourcing in which there is a requirement of some company or an organization to support the project in the beginning or the project has the specific shareholder to fund the project but in this type the dependency from the other people or organization is less as compared to reward-based crowdfunding.

A. DIFFERENCE BETWEEN CROWD SOURCING AND CROWDFUNDING:

Crowd funding and crowd sourcing looks exactly similar, but the main difference is that crowd funding is the specific type of crowdsourcing where only money or funds collected from the people i-e. to support the project finically while crowdsourcing is anything from the people whether it could be ideas or anything else or even funds. The examples of crowdsourcing include Wikipedia in which people from different countries are trying to share information although they are unknown to each other. So, Wikipedia is the crowdsourcing example where the information is getting collected from crowd. Similarly, Kickstarter and Indiegogo are the main sources to collect money from the crowd. So, these are the examples of crowdfunding.

As the rise of internet, Crowdfunding's demand is plummeted due to the availability of internet and access from across the world. The Online platform is an idea way to collect money or funds for the specific project is the best way to do it because it is available and easy to reach, and it is cheaper to make the website or an application and start crowd funding This thing also helps the people and Organizations to fund the Project which suits their Interest.

B. ADVANTAGES OF CROWD FUNDING:

The main advantage of crowd funding include it is one of the speediest ways of promoting your idea and be funded without the major investment from the pocket, People will fund the project based on their interest if it suits them and if the project gain popularity then many organizations and people will participate in the project and support them financially. Crowd funding also helps the people including the investors to know about view from people to understand whether to invest more or not in the specific project.

C. BUSINESS SUCCESS:

Kickstarter is one of the successful organizations which was founded in the year 2009 and this organization has received or pledges more than 4 billion of US dollars to fund more than 200,000 projects which includes different categories like games, music, movies, arts, poetry, technology, food and other different categories. The main objectives of this organization are to make innovative and promoting ideas to fund the project and taking money from different people and organizations to change the idea to reality.

II. MOTIVATION

While there has been work on the Crowdfunding dynamics, the work is still in its early stages. This project will work on the fact that there is still opportunity for getting new perspective in knowing project success in this perspective. By using data science tools and machine learning techniques, the project wishes to find unique results which are yet to be referred in the Crowdfunding literature.

III. OBJECTIVES

In addition to revisiting the already discussed results in the existing literature, the project looks to address the following questions:

• Do any of the category is important that is successful or not and is this important than other or not. So, the categories which are important there could be pattern which can be find later using machine learning algorithms.

• How did the success rate evolve over the years across different product categories? The project wants to test whether past success of a project has effect on the future success.

There are different machine learning algorithms could be used in this type of Project as this is related to business Intelligence to find the categories which are more important to make the project successful. More specifically, the problem is classification. So, the classification algorithms which include Support Vector Machines, Naive Bayes Rule, gradient boosting Machines, random forest, decision trees, logistic regression, KNN and Artificial Neural Networks and others. At the end of the project, there will be a comparison of different type of algorithm which is more accurate and useful for this specific type of application.

IV. DATASET

As for this project, there are many attributes in the dataset which includes ID, name, category, country, details of the state whether the project is successful or not, project launch, deadline date, funding goal. There are more than 300,000 rows and around 15 features in this dataset. The dataset is obtained from the very famous website Kaggle.
The categories of the project include games, film, poetry, music, restaurant and movies or videos
The data is scrapped from the Kickstarter Project and put in the csv file by the web scrapper in the year 2018.

V. LITERATURE REVIEW

Crowdfunding is a relatively new phenomenon, so it is no surprise that the related literature is only emerging. Here, we provide parallels with other sources of entrepreneurial finance to better understand the specificities of crowdfunding as a distinct form of finance. Taking crowdfunding from a purely financial perspective, we can make connections with bootstrap finance. Several studies provide evidence of the different forms of internal sources that bootstrapping entrepreneurs use [3]. (cosh et al. 2009) analyze other financing methods for start-ups and examine a broad range of financing alternatives [4]. (Agrawal et al 2011) focus on crowdfunding more specifically. They examine the geographic origin of consumers who invest on the SellaBand platform and observe that "the average distance between artist-entrepreneurs and investors is about 3000 miles, suggesting a reduced role for spatial proximity" [5]. (Mollick 2014) also examines the geography of crowdfunding using data from Kickstarter to study factors of success in crowdfunding platforms. Mollick discovers "a robust topographical element to the type of projects, with creators suggesting projects that attract the principal ethnic products of its topographical region" [2]. (Kuppuswamy and Bayus 2013) also examine funded projects listed on Kickstarter and show that social information plays a key role in the success of a project [6]. Ahlers et al. focuses on the data which is between the entrepreneur and crowd. They did analysis on equity crowdfunding which shows that credible information, quality of the beginning and sound info exposes to the crowd if the crowdfunding is successful. They achieved by using Australian Data [7]. (Nocke et al 2011) link product pre-ordering to price discrimination in a context of information asymmetry. In this case, the true quality of the product is revealed later, so the firm must deal with consumers with different expected valuations of its forthcoming product. One main difference with this literature is that crowdfunding requires that first period profits be above some minimum level [8].

Kickstarter is an American firm which has headquarters in New York, that has name in crowdfunding platform and their main goal is to do the project with the help of backers. Kickstarter got more than $1.9 billion of US dollars in pledges from 9.4 million dollars for funding more than 250,000 innovative projects. They target different categories like technology, games, journalism, poetry and others. The main people who are funding the Kickstarter projects presented very good rewards or practices in return for their effort to pledge the amount. Kickstarter is one of several crowdfunding platforms for taking money from the public. Project inventors select a target and a lowest funding goal. So, they canceled the project if the amount of money is equal to the funds must be collected within certain time period. There are many countries where the project, backers come from like United States of America, United Kingdom and others. In Kickstarter project, no one is owner over the project and contribution. The platform is easily reachable for everyone as it is web based. If the funding is complete, then they hold the information on the web which cannot be altered. Creators categorize their projects into one of 15 main categories. Some of the main categories include Comics, Dance Film, Video, Food, Games, Music, Photography, Technology etc.

Danhoung Chen discussed and focused on Young Entrepreneurs and proposed an idea of entrepreneurship crowdfunding website which is good and support the college student entrepreneurs and get rid of the problem of old and existing entrepreneurship methods. So, this thing will fill the gap between young entrepreneurs and crowdfunding and provide a very good in terms of enhancing business, entrepreneurship and increase crowd funding for the specific project using the crowdfunding website in China. [9].

Lester Allan Lasardo et.al uses dataset rather than the most common and famous Crowdfunding platforms like Kickstarter and Indegego, uses the dataset from Crowdfunding source which is working only in Finland instead of working globally. They study the data which is focused on the main cities in Finland which include Helsinki and Tampere region using reward based and equity Crowdfunding to get the attention of young entrepreneurs in Finland by following the successful crowdfunding sources like Kickstarter as of inspiration. In their research they compared the reward-based crowdfunding as well as equity-based crowdfunding based on success rate in Finland. They found out that both types include reward-based crowdfunding as well equity-based crowdfunding are important for crowdfunding in Finland. Using successful campaigns, equity-based crowd funding has upper hand in Finland due to the most the funds gained in this category while in Helsinki, the popularity is equal to due to large population and most of the population of Finland in Helsinki [10].

Liu Xuefeng et al tried to suggest a unique business model of crowdfunding platform centered on the explanation and evaluations of many improvements created by a Chinese platform enterprise involving in crowdfunding for hostels etc. In this model, company merges reward-based crowdfunding with equity-based crowdfunding. This type of crowdfunding models made the modifications where people who fund the project represent as customers as well as stockholders at once. In this research, they also proposed a Chinese crowdfunding platform which supports the people who raised the voice for funding and the people who support them in terms of filling the gap between them. So, if there is a connection between them then it supports crowdfunding and project.

In their proposed method which is the creation of the business ecosystem which merges all the organization which includes online companies handles business like ticket system and investment firms, entrepreneurship service, smart hotel management, content service bolsters the source of information which in turns help crowdfunding to the next level [11].

.Rafat M. Abushaban take deeper look into the term crowdfunding by investigating it and how crowdfunding effect on Projects and businesses which are new for the crowdfunding development specifically in Middle East.

This paper also proposed how to increase crowdfunding and possibilities for campaigns in Middle East by removing the barrier around it. They scrapped data from Indiegogo based on the reason that it operates and do research in European and north American countries but the Indiegogo focus on almost every country on the world platform and analyzed the projects which are based on success including partially successful and failed project as well. In the end of this research they visualize which category is more successful and gained the maximum amount of funds that is Education and Film. and, which country in the middle east has gained more attention to the crowdfunding that is Palestine [12].

Meilin shi and Lei Guan made the dataset of Technology Projects from Jing Dong crowdfunding platform and studied deeply on the factors which contributing the successful crowdfunding. They used technology project and used Regression analysis on dataset, and they found out for making project successful which includes video is more helpful for raising funds instead of images as people believes more in video [13].

Zhanyang song et al. also uses similar approach like [9] and suggests young entrepreneurs to achieve their goal by using crowdfunding platform which is for entrepreneurs to fulfil the needs of young entrepreneurs. [14].

Yusan Lin et al analyzed reward based crowdfunding platforms by collecting data from Kickstarter and focus on the success of the project. They used many data mining techniques to find the specific pattern which includes Kolmogorov-Smirnow test and Kaplan-Meier estimation. In this research paper they classified into project which is going to be successful using these techniques. [15].

Kevin Chen et al. made an android application as well as chrome extension to predict the project whether it is successful or not with 67 percent accuracy. The algorithm they used behind for the prediction is support vector machines as it is one of the most important and common algorithms for supervised classification type of problems. The data they used for prediction is from Kickstarter by scrapping it and worked on that data [16].

Zhiyuan Tian et al. has done a research by finding the important factors of successful Internet crowdfunding projects. The two most important platforms they targeted include JingDong crowdfunding platform and TaoBao crowdfunding platform which only doing progress in China. They used logistic regression as a classification algorithm to predict the project success. Also, they found out every platform has different effect on success of the project and video has positive impact on the success at a significant level of 10 percent [17].

Francisco Ferreira and Leandro Pereira analyzed the pattern of motivation behind the reward-based crowd funding and equity-based crowd funding and found out motivations for investors in the equity model and reward based looks same. So, based on their findings on reward baaed crowdfunding could be applicable to equity-based crowdfunding which lacks data [18]. Peng Du et al. focused on reward-based crowdfunding when the backers got their profit in terms of rewards which step is mandatory for crowdfunding to manage the funds gained by crowdfunding which is known as post-crowdfunding. There are two stages which include in this term post crowd-funding are resale sale and sale which is usual and normal. They developed a two-stage model for taking a closer look into the resale policy, strategic purchasing behavior. For price prediction they used genetic algorithm along with backward induction based on optimization decision making. They also explored managerial intuitions using numerical case study [19].

Jeya Gera and Harmeet Kaur focused on textual analysis from text, profile photo and details of campaign on project from the Kickstarter project. With the help of tools available for textual analysis they found the correlation between campaign achievement and personality behaviors, so they focused on overall predictive value of personality even with low success. They classified into Openness and Extraversion based on achievement on campaign using Logistic Regression [20].
Yan Li et al. they focused on success of the project from Kickstarter along with tweets from Twitter and used Logistic regression and Log-Logistic Regression for finding the pattern in regression analysis. They also analyzed factors contributing the success on the Project [21].

VI. COMPARATIVE ANALYSIS

In this section we compare the most important methods and could be applicable to our dataset in a very brief view.

So, the first column indicates the paper number in the reference section while second column shows year of the publication of the paper while the third column indicates type of the crowd funding focused on the paper. Similarly, the last two column indicates the source of Dataset and Brief Findings in the paper respectively [2].

No.	Year	Type	Method	Data set	Finding & Result
[15]	2016	Reward Based Crowd Funding	Kolmogorov-Smirnov test & Kaplan- Meier test	Kick Starter Project From 15/12/2013 to 23/03/2014	Focused mainly on success of project found on the dataset. Project with more rewards, limited editions achieve more goals of the project pledged.
[16]	2013	Reward Based Crowd Funding	Support Vector Machines	Twitter & Kickstarter Project Website	Achieved 67 percent accurate result on crowdfunding project success
[17]	2018	Reward based Crowd Funding	Logistic Regression	JingDong crowdfunding platform and TaoBao crowdfunding platform.	Quality signal include Goal, Period, Video are significantly important similarly SLP and FLP has bad impact on the prediction of successful project in project KIA decrease support for backer while AON increase support for backers
[19]	2017	Post Crowd Funding In Reward Based Crowd-Funding	Genetic Algorithm & Backward Induction	Focuses on Improving Business model	Two stage model is developed in closed analytical form. Classify the optimal Price in the Normal Stage while Genetic Algorithm is used For having solution of resale Price.
[20]	2018	Reward Based Crowd Funding	Logistic Regression	Kickstarter & Kickspy	Focus is on the personality traits on funding From the Backers. Prediction Model worked with 75 Percent accuracy using Logistic Regression By Investigate family traits.
[21]	2016	Reward Based Crowd Funding	Logistic Regression Log-Logistic regression	Kick Starter Project Website & Twitter	Achieved good accuracy in prediction of algorithms When used with failed projects as well. They scrapped data for 3 days and the accuracy is improved. The highest accuracy achieved with log-logistic regression about 0.90 area Under the curve.
[09]	2016	Equity Crowd Funding	Proposed Solution	NA	Proposed in improving Existing crowdfunding in young Entrepreneurs in China by developing university student entrepreneurs tool using TRSADMIN
[10]	2014	Equity Crowd Funding	Data Visualization only	Crowd Funding source working in Finland	Based on different cities equity crowd funding has upper hand in Finland with the help of visualizations.
[12]	2014	Reward based Crowd Funding	Data Visualization only	Indiegogo platform	Data Visualization of different countries specifically in middle east about the success and failure contribution from the scrapped data.
[13]	2016	Reward based crowd funding	Linear Regression	Jing Dong Crowd funding platform	Finding category based on linear regression to get the project successful in the dataset

VII. PROBLEM STATEMENT

The problem with the project is that without getting the details of the dataset we cannot say what type of problem we are going to solve. With the help of visualization of techniques using Logistic Regression, we can infer that that categories we assumed are good for the prediction of classification based on different features. First, we tried to predict the classification based on the features. In most of the research, logistic regression is the most common techniques used in different research papers. There are many categories used for crowd funding and the data visualization of these categories with respect to the state of the project which include cancelled, failed, live, successful, and undefined. Figure.1 showing the comparison of the main categories of the project with respect to the state. In the x-axis, which is representing the number of appearances of category in the dataset w.r.t to the state in different colors and y-axis representing the category. Film and Video is the category which appear the most as compared to others but the failed contribution in this category is higher as compared to successful projects. Similarly, Music is the second most appeared category in the dataset and the successful projects are more as compared to failed and other states in this category.

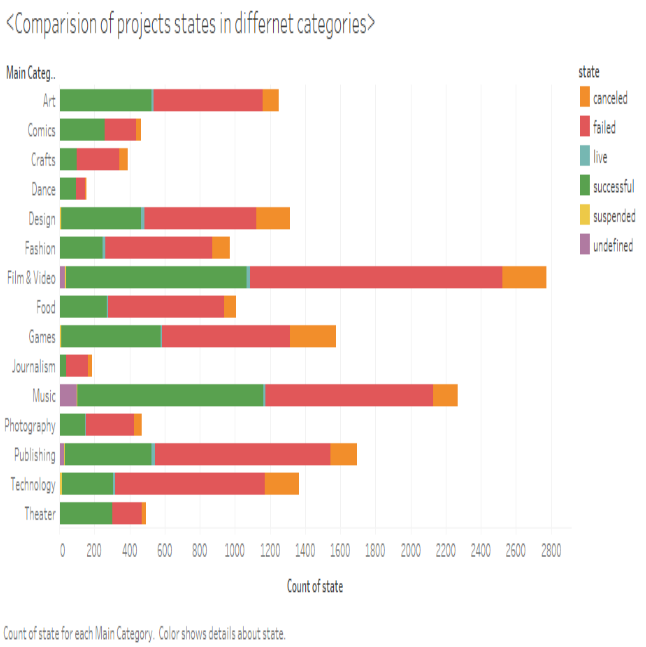


Figure.1: Showing the visualization of project states vs categories

So, with the help of visualization using Tableau indicates that most of the projects either successful or failed that's what in other people work.

But most importantly, there is a significant amount of appeared categories which are ignored by almost everyone that include cancelled and live projects. So, that's the research gap.

With the help of features used for the prediction of cancelled and failed projects are categories, Goal: the amount of money in US dollars. Similarly, amount of money in US dollars pledged.

In the research, Logistic Regression has reached with the accuracy of 89 percent as we tried it for cross reference, and we got the same result with including the features. Similarly, we tried gradient boosting machines and got the accuracy 97 percent.

The Problem include in this project is to apply different machine learning algorithms like logistic regression, gradient boosting algorithm and deep learning machine learning artificial neural networks with forward feed to check the accuracy in the multi-classification problem.

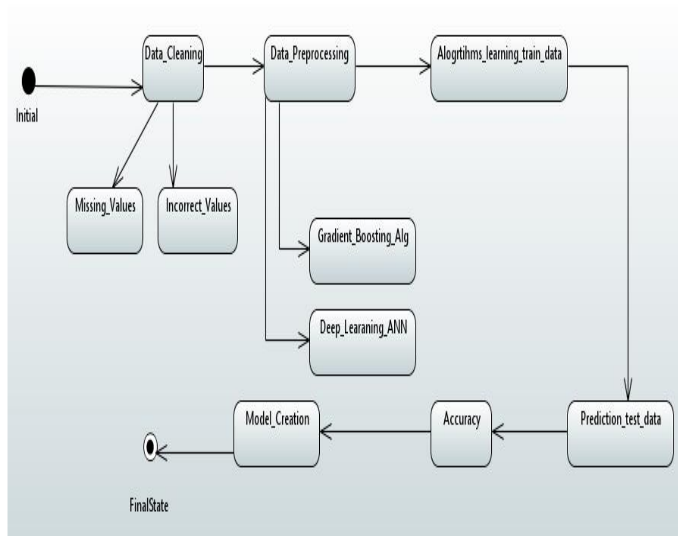


Figure.2: Showing the complete process of Project

The Figure.2 showing the complete process of the project which includes many stages. The first stage which include the Data cleaning has already been done for the prediction of success and failure to confirm the related work with gradient boosting machines and logistic regression.

The second stage of the Data Analysis is Data Pre-processing which has done before based on binomial classification. So, we cannot use logistic regression for this type of specific multi classification problem, So, we will use Deep Learning Artificial Neural Networks along with Gradient Boosting Machines to predict the classes whether the project is successful, failed, live or cancelled.

Table.1 showing the number of projects appeared and their percentages. If look more closely, it can be saying that failed projects appeared more as compared to others. Followed by successful, Canceled, Live and Suspended

Status	Number of Projects	Percentage
Suspended	1842	0.49%
Live	2798	0.75%
Cancelled	38751	10.34%
Successful	133851	35.71%
Failed	197611	52.72%
Undefined	0	0.00%

Table.1: Showing the significance of status appeared

People in the previous work did not consider the other categories of the state of the project which includes live and cancelled and others. As the existed work only focused on the Failed and Successful project but we take a closer look at Table.1 there is also a significant number of Projects which are cancelled as shown in the table.1 which constitutes around 10 percent in the dataset.

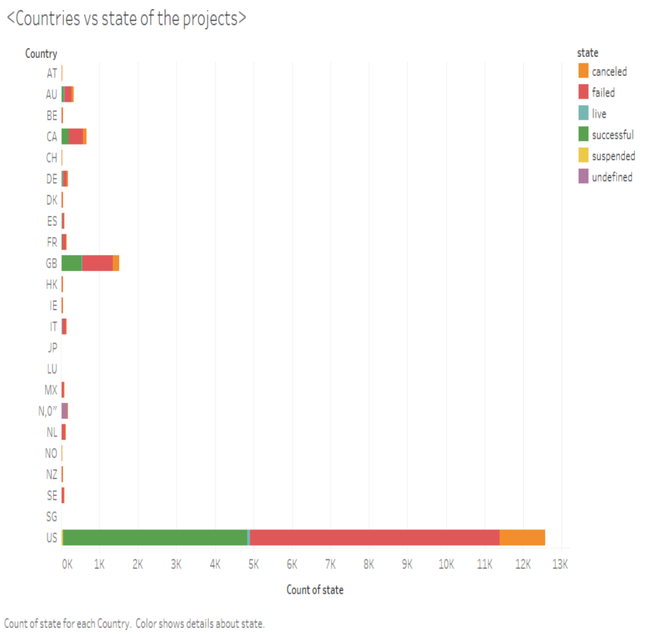


Figure.3: Showing the frequency of projects occurred in different countries.

The most important country is USA as it has the highest amount of pledged amount come from followed by Great Britain, Canada and Australia. The x-axis representing counts of appearance of the countries while y-axis representing the country names. The reason for the appearance of the US is due to the fact as the Kickstarter is the US firm and most of the data scrapped from the Kickstarter focused on North America.

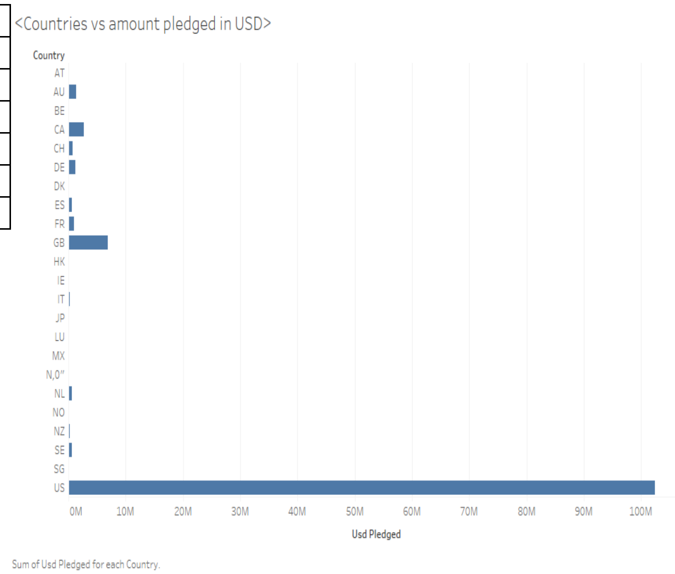


Figure.4: Showing USD Pledged from different countries

Similarly, Figure.4 showing the amount pledged in USD in different countries. The reason for the appearance of funds from US is due to the fact US appeared the most which makes sense to be funds from the same country. So, in the x-axis of the figure which is representing the amount of funds in USD which is showing around 100 million of USD come from US

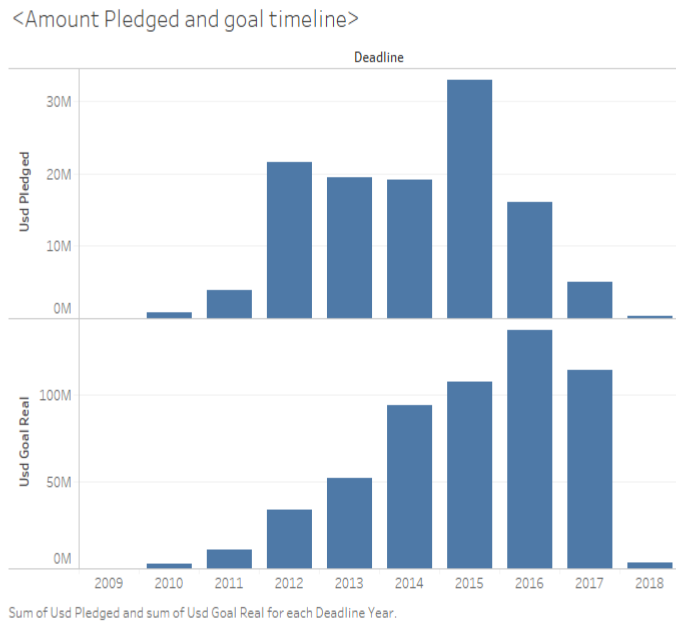


Figure.5: Showing the amount and goal timeline

So, the figure.5 showing the amount and goal in different years. The amount pledged with highest amount in the year 2015 with more than 30 million of US dollars similarly, the goal amount reached more than 100 Million of US dollar peak value in the year 2016.

As the Kikcstarter founded in the year 2009. So that time it was a start, so no amount pledged and goal as there are no projects. As the time passes, new projects occurred similarly the amount pledged, and goal money increases significantly. According to their website they declared their organization to give benefits and favors to the people by completing the good ideas to reality. In the year 2015, they made this announcement publicly that their organization is dedicated to the public and working with the help of people in the form of backers. So, the reason is increasing the peak amount of US dollars in pledged as you can see in the figure.5 showing the highest amount of 300 million of US dollars in pledged amount.

VIII. PROPOSED SOLUTION

A. DATASET DETALIED ANALYSIS

In this section, the more detailed view of the problem and its proposed solution explained.

The dataset which is collected from Kaggle has many features which are shown in the table. II

S.No	Feature	Type
1	ID	numeric
2	Name	string
3	Main_cateogry	factor
4	Currency	numeric
5	Deadline	integer
6	Goal	numeric
7	Launched	integer
8	Pledged	numeric
9	State	integer
10	Backers	numeric
11	Country	integer
12	USD pledged real	numeric
13	USD goal real	numeric
14	USD pledged	numeric
15	category	Integer

Table. II shows the Initial features in the dataset

There are 15 features in the dataset. ID represent the specific number for the project occurred in the Kickstarter Platform. Name represent the project name in the crowdfunding platform. Main category represents the main category of the project funded. Currency represent the funds collected in different currency while Country represent the detail of the backer's and their country. Deadline and launched represent the duration period of the project. Goal amount is the amount needed to make the project successful while pledged is the amount pledged by the backers and USD pledged real is the amount of pledged in US dollars. USD pledged is like USD pledged real, so it behaves as a noise and similarly, USD goal shows the amount of money needed to make the project successful in USD. Category represent sub category of the main category. State is the dependent variable of the dataset.

ID feature has no importance in the dataset because ID has no significance. Similarly name of the project does not matter as it is not related. USD pledged column shows similar numbers compared to USD pledged real, so an assumption has taken to consider it as noise. In the category section, there are more than 60 categories which include Product Design, Documentary, music and video games etc. which are subset of the main category's column, so here to make it more simplified, category column also considered as noise.

Currency column represents the country as the country column is also there in the dataset. So, currency column is also removed from the dataset. Launched column showing the date of the project launched while the Deadline represents the date of the project to finish, after that date no money asked. So, for data analysis in launched column minutes and hours has been removed to make it like the pattern in deadline and then use as. POSIXct function in R programming to manipulate these two columns into one date_diff which represents the difference in days to put into an algorithm so algorithm might understand the pattern.

So, in the end, following features are left for data analysis as shown in the Table. III.

S.No	Features	Type
1	Date_diff	Numeric
2	USD_Pledged_real	Numeric
3	USD_Goal_real	Numeric
4	Backers	Numeric
5	Main Category	Factor (Dummy Required)
6	Country	Factor (Dummy Required)
7	State	Factor

Table.III shows the extracted feature

State variable which is the dependent variable. It has six features which include the success, failed, canceled, live, undefined, and suspended.

In the previous literature review and research, authors only focused on the Failed and Success in the state variable, but other categories not be considered.

So, in this paper four categories are going to be considered based on importance which include failed, success, canceled and live. Undefined and suspended has been removed from the dataset and considered as noise. So, A number is allocated to predict the class. There are 4 classes in which 0 represents failed, 1 represents successful, 2 represents live and 4 represents canceled.

There are 21 countries in the country column, so there are 21 dummy variables are created to represent separate column for each category because algorithm considered as numeric factor where higher number represent higher value while smaller number represent smaller value. US is the most dominate category in the column.

Similarly, the dummy column also has been created for the main category separate for each category. Film and video, games, publishing and music are the top categories which appeared in the dataset.

Column backers represents how many organizations are backing the project.

The missing values have been removed as there are insignificant amount of missing values represents less than 5 percent of the whole dataset.

B. DETAILED DATA ANALYSIS

The correlation between the factors are shown in the table. IV as well as in the figure.6 using package corrplot in R.

	State	Number of backers	USD Pledged	Goal	Date diff
state	1	0.034	0.0319	-0.003	0.008
backers	0.037	1	0.75	0.004	-0.08
Pledged	0.03	0.453	1	0.005	0.001
Goal	-0.02	0.004	0.005	1	0.004
DD	0.008	-0.0008	0.001	0.004	1

Table. IV Showing the correlation between continuous variables

In the Table. IV which shows the continuous variable in which pledged and goal in USD while DD represent date difference in days. If we compare the correlation between state. By observing the table, it can be concluded most of the variables with respect to the dependent variable state has almost zero correlation except the backers and USD pledged which has correlation. As shown in the figure.6

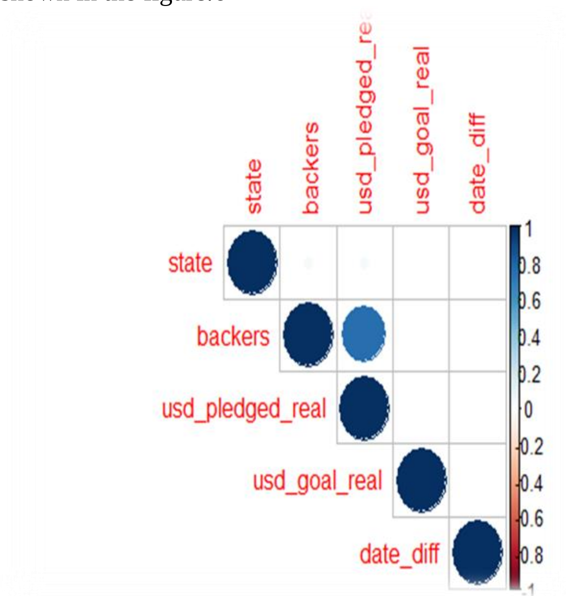


Figure.6: Showing the correlation using corrplot.

The features we have here are number of backers to fund the project, the amount of money pledged by the backers to fund the project, Date difference in days, The target amount and the countries of the bacekers support the project. The main details and description already explained in the earlier section.

The features which shows less correlation linearly in between them. To determine the more details of the features, linear Discriminant Analysis. There is a library MASS in R used to evaluate LDA analysis of the continuous features which are backers, pledged amount in USD, goal amount in USD, date difference in days compare to continuous variable state.

	backers	Pledged	goal	DD
Failed	7.4e-5	6.4e-5	3.8e-4	0.002
Success	1.2e-4	0.001	5.73e-5	0.002
Live	3.1e-4	0.0002	3.75e-4	0.002
Canceled	1.1e-3	0.0001	4.54e-4	0.002

Table V. Showing the means of continuous variable using LDA

So, the values in table. V represents the means for each covariate, while in table. VI shows the linear combination coefficients also called scaling for each linear discriminant. Linear discriminants are based on the number of classes we have in the dataset. It is always less than one at most from the number of classes in the dataset. In our dataset, there are 4 classes and there are 3 linear discriminants as shown in the table VI.

	LD1	LD2	LD3
Backers	187.5	30.12	-69.37
Pledged	53.09	29.68	105.93
Goal	-26.14	25.11	134.99
DD	-51.22	234.45	-53.38

Table VI. Showing the scaling values using LDA

Similarly, by retaining the result from LDA function result, singular values for the LDA1, LDA2 and LDA3 are 48.39, 6.04 and 0.33 respectively. Singular values interpret the ratio of the between and within-group standard deviations on the linear discriminant variables.

Singular values could be utilized to calculate the variance between the group which can be understand with the help of Linear discriminant. In our business problem it could be seen that the first linear discriminant explains more than 98 percent of the between group variance in the Kickstarter dataset.

LDA1:0.98
LDA2:0.0015
LDA3:4.59e-5

Similarly, LDA2 is around zero percent of the between the group variance in the dataset and same thing with the LDA3.

C. ARTIFICIAL NEURAL NETWORK PREDICTION MODEL

Artificial Neural Networks (ANN) systems are intelligent computing systems that resemble the biological neural networks in human brains. An ANN consists the networks of connected units or nodes called artificial neurons. In an ANN, a typical artificial neuron receives the signal, process it according to activation function used to program it to send to the next artificial neurons connected to it via a connection between two consecutive nodes.

One of the most popular ANN paradigms is the feed-forward neural network (FNN) and the associated back-propagation (BP) training algorithm. Feedforward Neural Networks are the type of artificial neural networks where the connections between do not form a cycle. Feedforward neural networks were the first type of artificial neural network invented and are simpler than their counterpart, recurrent neural networks. They are called feedforward because information only travels forward in the network (no loops), first through the input nodes, then through the hidden nodes (if present), and finally through the output nodes [22]

In our modified Artificial Neural Network,

Input Layer: This is the initial layer of a neural network supply the input data or features to the network. We have 40 layers of input which including 37 dummy variables and 3 continuous variables

Output Layer: This is the final layer which gives out the predictions. As we have 4 output layers which shows the state to be accurately identified using ANN as we have four outputs

Hidden layer: A feedforward network applies a series of functions to the input. By having multiple hidden layers, we can compute complex functions by cascading simpler functions. The number of hidden layers is termed as the depth of the neural network. Hidden Layers selected based on the research [23], as they used two hidden layers.

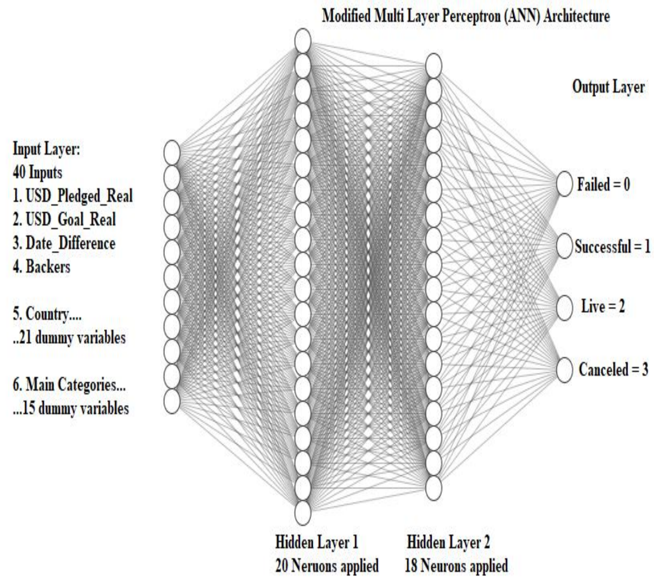
Activation Functions:

The activation functions used in

- Softmax
- Rectified Linear Unit (ReLU)

There are many other activation functions as an activation function but in our problem, rectifier linear units are used to save from the vanishing gradient problem if used sigmoid or tanh. Softmax used in the output layer to extend our model from binary classification to multi classification.

D. IMPLEMENTATION DETAILS



.

Figure.7: Showing the architecture of our ANN Model

The input layers are 40 which include the backers, USD_pledged_real, USD_goal_real, date_diff, 15 dummy variables of main_category and 21 dummy variables of country which constitute 40 input variable feed into the neural network. The output layer has four outcomes from 0 to 3 where 0,1,2,3 represent failed, successful, and live and canceled respectively. The tool "KERAS" is used for the implementation of artificial neural network which is the python library for computations of deep learning calculation. Keras uses TensorFlow on its backend. The implementation using Keras is simple and straight forward.

Input Layer	Hidden Layer 1	Hidden Layer 2	Output Layer
Paramters	Neurons	Neurons	Parmeters
USD pledged amount	20	18	Failed
USD Project goal amount	NA	NA	Succesful
Date Difference between start and end	NA	NA	Live
No. of Backers support the project	NA	NA	Canceled
Country US	NA	NA	NA
Country_SE	NA	NA	NA
Country_NZ	NA	NA	NA
Country_NO	NA	NA	NA
Country_NL	NA	NA	NA
Country_ML	NA	NA	NA
Country_MX	NA	NA	NA
Country_LU	NA	NA	NA

Country_JP	NA	NA	NA
Country_IT	NA	NA	NA
Country_IE	NA	NA	NA
Country_HK	NA	NA	NA
Country_GB	NA	NA	NA
Country_FR	NA	NA	NA
Country_ES	NA	NA	NA
Country_DK	NA	NA	NA
Country_DE	NA	NA	NA
Country_CH	NA	NA	NA
Country_CA	NA	NA	NA
Country_BE	NA	NA	NA
Country_AU	NA	NA	NA
Country_AT	NA	NA	NA
Cateogry:Art	NA	NA	NA
Cateogry Comics	NA	NA	NA
Cateogry Crafts	NA	NA	NA
Cateogry Dance	NA	NA	NA
Cateogry Design	NA	NA	NA
Cateogry Fashion	NA	NA	NA
Cateogry Film & Video	NA	NA	NA
Cateogry Food	NA	NA	NA
Cateogry Games	NA	NA	NA
Cateogry Journalsim	NA	NA	NA
Cateogry: Music	NA	NA	NA
Cateogry Photography	NA	NA	NA
Cateogry Publishing	NA	NA	NA
Cateogry Technology	NA	NA	NA
Cateogry Theater	NA	NA	NA
Activation Function	**Activation Function**	**Activation Function**	**Activation Function**
Rectifier Linear Unit	Rectifier Linear Unit	Rectifier Linear Unit	Softmax
Loss Function		Categorical Entropy	
Optimizer		Adam	
Batch Size		64	
Epoch		100	
Metrics		Accuracy	

Table.VII: Shows the implementation details of our ANN Model

As shown in the table. VII which is explaining all the implementation of details of neural network in Keras. The reasons for selecting Rectifier linear unit as an activation function in the input and hidden layers because there is a problem of vanishing gradient which occurs in hidden layers, to confront this problem, rectifier linear units are used. SoftMax activation function is used in output layer because of problem is multi-classification.

Similarly, in compiling the model, due to classification, cross entropy is used as a loss function. For optimization, Adam is used while 100 number of epochs along with 64 batch size used in the implementation of Artificial Neural Network as shown in the figure.9

We increased the number of neurons to see that whether to increase accuracy, but the accuracy is decreased. Increasing is not always gives better result sometimes it worked the opposite. Similarly, the hidden layers but increasing hidden layers mostly increase the accuracy unless it saturated to the specific accuracy limit. Same thing with epochs and batch size, these are the most important parameter for tuning in terms of running neural network for different times to check which combination gives the most optimum result without overfitting, under fitting or gradient exploding and gradient vanishing problems.

IX. EVALUATION

In this section, we discussed the evaluation of our findings in this paper. The previous section discussed the method details which include the implementation issues and architecture of our model.

A. PARAMETER TUNING

The most important, difficult and complex step in using the neural network is the parameter tuning which takes a lot of time and requires a lot of combinations and testing to get the best result based on the accuracy of the model.

So, without proper tuning model could work pretty based on the choose of the wrong parameters which is always different based on the type, nature and complexity of the dataset. In our model, there are certain parameters which cannot be changes due to the nature of the problem. These parameters include the shape of the input layer and output layers. In our case it will remain 41 and output layer shape is 4 because we are classifying multi class of success, failure, live and canceled projects. Evaluation metric is also remaining accuracy, Loss function is also unchanged as the classification is multi so we used categorical cross entropy, but it could be change based on the application for example in the previous research people used for binary classification as binary cross entropy.

The parameters which required tuning in our model include the number of layers, number of neurons, number of hidden layers, neurons applied at every layer, batch size, number of epochs, activation function at every layer. We first started with one hidden layer then we increased it to check the effect of layers on the accuracy of the model. We stopped at the two hidden layers because if we increase more layers, we faced the problem of vanishing gradient as the model is not even predicting the category in the output.

In the next step, we target the neurons applied at each layer along with the activation function. For the input layers we used rectifier linear unit to keep the model away from the vanishing gradient problem. In the last layer, SoftMax is used as it is good for the prediction of multi class. Similarly, we come up with neurons 100, 20 ,18 on from input layers to two hidden layers and gives the good accuracy as compared to other combinations.

In the next step, we tried with different number of Epochs and batch size to check the model accuracy. First, we take very small number of epochs then we increase the number of epochs till 100 epochs and even increasing more does not affect model accuracy, so we ended up in having 100 epochs in our model. Similarly, for the batch size, we started with the 32 and ended up in 64 by applying different combinations of taking numerous batch size.

Figure.9 shows the number of epochs on its x-axis along with the loss function on the top graph while on the bottom, it is showing the accuracy. We can observe the loss functions was decreased as the number of epochs and reached to 0.50 while on the same time accuracy is increased. The green line is showing the training while blue line shows the validation findings.

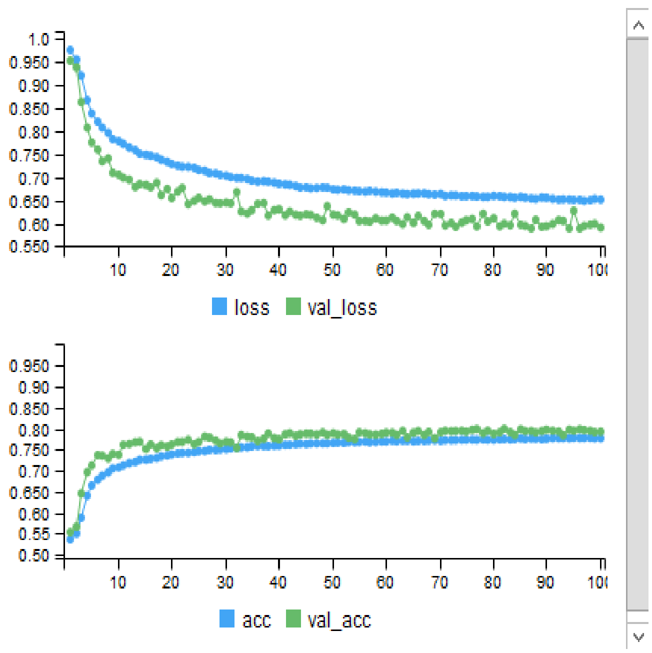


Figure.9 showing the accuracy and loss function status

B. EXPERIMETNAL RESULTS

We discussed the evaluation of our model which is based on the most important factor accuracy of an algorithm.

Training set	**Test set**
298488 observations	74518 observations
80 percent dataset	20 Percent dataset
81.6 percent Accurate Result	79.7 percent Accurate result

Table VIII.: showing the accuracy of the model

We include the whole dataset for the evaluation of our methodology. table. VIII showing the accuracy of the model. We considered the whole dataset which has more than 300,000 observations for the whole data. We divide the dataset into train and test data which has around 300,000 thousand observations and 75,000 observations respectively which represent 80 percent and 20 percent of the data respectively.

C. COMPLEXITY AND EFFICENCY ANALYSIS

If we considered the complexity of the algorithm as it is based on the smaller number of neurons and less hidden layers as it is usually taken for multi classification to have the better accuracy.

The model has more than has 298,488 observations for the train set, so it takes more time to run the model. We used local machine of 8 GB of main memory and used Intel Core i5 of 2.40 GHz with windows .10. As the system is not efficient of the computation of numerical processing that is why it takes 30 minutes to run the process.

In terms of efficiency, we have around 80 percent of the result is accurate which means out of 100 observations, 80 could be predicted accurately.
If we go more into the detail of the algorithm used based on space complexity and Big O Notation.

EPOCHS	**BATCH SIZE**	**NUMBER OF ITERATIONS**	**TRAINING ROWS**
1	64	4664	298488
100	same	466400	same
Total Parameters Used for training: 40,062			

Table. IX: Showing State complexity of the method

Table. IX showing the total number of iterations required to train our model so the total iterations are 466,000 to complete 100 number of epochs which takes 30 minutes to run the model in our local machine.
If we see the Big O notation of the optimizer, we, used Adam optimizer which is the optimized method of stochastic gradient boosting algorithm but almost every optimizer relies on back propagation learning method. If we consider n number of layers, n number of neurons, and representing node multiplication with n. then our complexity of algorithm becomes [24]. **O (n^k) So**, the general formula for representing the time complexity of our algorithm is

O (nt*(Matrix Multiplication) --------------(1) [25]

where n = number of epochs, t represents the trains samples

D. COMPARING WITH OTHER ALOGRITHM

As we diagnosed in our main method which include the problem of reduce variance and bias and overfitting. Although we handled the overfitting by doing scaling and parameter tuning but the problem of variance remained as we checked through linear discriminant analysis in the section. VIII.

So, to handle this we will try to think to use ensemble learning algorithm which is the term uses different algorithm to obtain good accuracy.

Instead of analyzing the whole dataset, we take the down sampled data based on the same proportion of the dependent variable state. To accomplish this, we used to create data partition function in caret to get the 10 percent of the data which represents the same proportion of occurrence of state as in the whole dataset.

The library used in the R for the gradient boosting machines algorithm is not efficient for handling the large data and large computation. We tried with whole dataset, but it takes a lot of time, so we abort the process. For this reason, we tried with down sampled dataset. It worked well; Even achieved more accuracy as compared to our main method.

Gradient Boosting Machines	
Ensemble Learning Algorithm	
Library	R
Distribution	Multi Nomial
Number of Trees	3000
n.minobsinnode	10
verbose	True
Shrinking	0.01
Interaction Depth	4
Train Data Used	

Table. X: Showing the gradient boosting machines parameters

Now, we discuss the parameter selection based on the type of our problem. Distribution used in the model is multinomial because of multi classification. We put verbose is TRUE to observe the parameters. We set the number of trees equal to 3000. Similarly, other parameters include the shrinking, interaction depth and minobssinnode in the method.
Table X showing the parameters used in the algorithm.

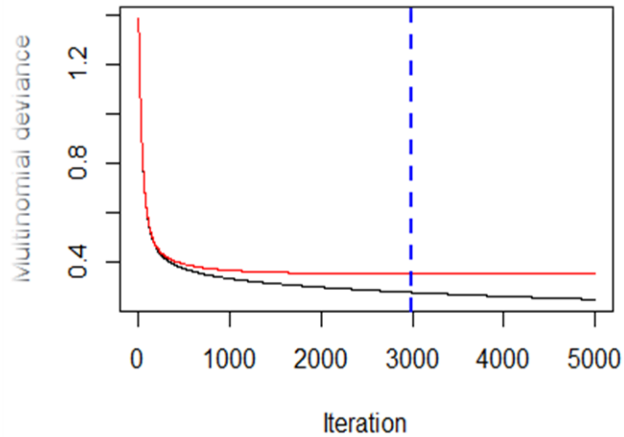


Figure.11: Showing the number of trees required

Figure.11 showing the how many numbers of trees required for the algorithm to converge. So, with the help of parameter tuning we reduced to number of trees required 6000 before to 3000 with the help of changing parameters. So gradient boosting machines takes 270 seconds to run which is less as compared to run time before tuning was 7 minutes.

Variable Names	Importance Based on GBM Method
Usd Pledged Real	37.13
Backers	34.06
Usd Goal Real	23.20
Date Difference	2.26
Main Category	2.15
Country	1.08

Table. XI: showing the importance of variables

Table. XI showing the relative influence of the variables. So, we have two categorical variables and 4 numerical variables. USD pledged real is the most significant variable followed by backers, USD goal real, date difference and the least are main category and country as they are categorical variable.

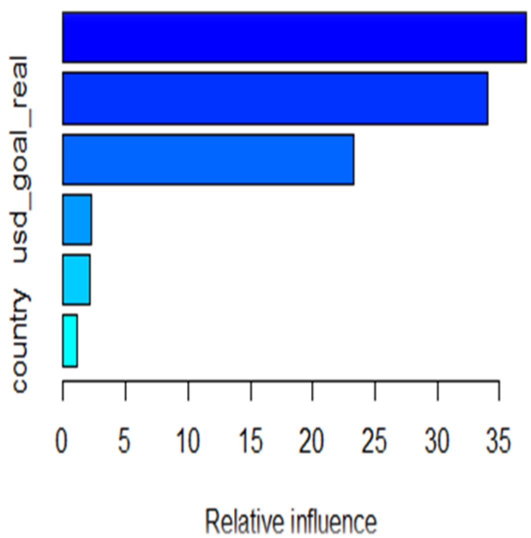


Figure.12: Showing the relative influence of the variables

In this figure.12 which is representing the table. IX in the graphical form showing the highest importance of USD pledged real which is around 37 while Country showing the least importance of 2.

Down Sampled Training Data Set	Down Sampled Test Data Set
80 percent of the down sampled data	20 percent of the down sampled dataset
30,015 observations	7377 observations
88.15 percent result	87.5 percent accurate result

Table. XII: Showing the accuracy of the GBM method in down sampled data

Table. XII which is showing for the gradient boosting machines to reduce overfitting and reduce variance problem. We have around 30,000 observations and around 7000 for the test set which is exactly the 10 percent of the down sampled data. We used same proportion of percentage that is 80 percent and 20 percent for the train and test data respectively.

The accuracy we achieved in this method is 88 percent when it is applied on the train data and achieved 87 percent when it is applied on the test data. So, if we take a closer look into the numbers, we did not have any much difference in test and train which implies that our method is not overfits and achieved good accuracy as compared to our main method i-e. modified artificial neural network.

Similarly, by checking the time complexity of this algorithm does not take more money as the functions used in gbm library is gbm which is not consist of any nested loops or loops, so we say the complexity of the algorithm is

O (logn)

E. COMPARING PREVIOUS WORK

In the previous work, most of the people focused on the success and failure contribution and visualization of the business idea of crowd funding. People scrapped different features which scrapped from Kickstarter.

Along with the previous literature review, there is significant work done most on different datasets, but one papers published on the same dataset which we have worked on.

Let’s discuss what they have done so far and what they missed. As discussed in the literature review section of this project, success and failure contribution only focused. Once the idea generated, people more on trying to do more efficient binary classification based on the accuracy of their implemented algorithms the most common method used is Logistic Regression and random forest for the binary classification.
Pi-Fen Yu [23]. used Kaggle dataset.

Model	Previous Work [23]
Binary Classification	**Accuracy**
MLP	0.93
Random Forest	0.929
AdaBoost	0.924
SVM	0.90
Decision Trees	0.90
Logistic Regression	0.89
Naïve Bayes	0.71
Multi Classification	**Our Work**
Multi-Layer Perceptron	0.80

Table. XIII: Showing the comparison of previous and our work

As they have achieved very good accuracy in terms of the prediction of the success and failure contribution in the Kickstarter dataset which is collected from Kaggle.

But the problem is that there are more than 2 categories in the dataset like Live, Canceled and suspended project which is considered as a noise in the previous research.

Here, in our project, although we achieved less accuracy as compared to the previous work but due to the different type of the problem, we can say that other method focused on the success and failed contribution, but we are focusing on other categories of live and canceled along with it.

The first paper who focused on the Kickstarter published in the year 2013 [16] which only considers the success contribution, but that time accuracy was too low achieved 67 percent. But as the concept become older people trying to focus more and more on the accuracy.

In the year 2013, [23] Pi-Fen Yu published using the deep learning multi-layer perceptron and achieved 93 percent accurate result on predicting success and failure contribution in the dataset as shown in the table. XIII

So, we achieved with 80 percent accuracy for the prediction of success, live, failure and canceled projects by opening a new area of research for the better accuracy.

X. CONCLUSION & FUTURE WORK

The main objective of this research project is to develop a model which could be used for the prediction of classifying different type of projects accurately and with the help of this we can used this model as a reference can apply on the unsupervised data to diagnose the problem and a pattern in the data of crowd funding and helpful to predict the type which should be put before the project started. This thing will also help the project to be successful. For example, if the model predicts that the project is going to be failed before collecting then it will save time, money and reputation of the platform. This will also help the other platforms of crowd funding to start a business and based on the model will help to observe the trend in the industry.

If we put the whole idea of the project in the nut shell, we applied data visualization techniques to understand the trend of categories of project. The visualization results incur that most important category to make your project successful is beneficial is film and video because it occurred the most in the dataset with respect to success followed by music. So, these are the categories to spend and start projects because projects success based on the public fund and interest.

Now we, moved to other details which say that most of the funds come from US. So, we must know what interesting topics and people are like the most to invest based on the current period.

The third thing is we used multi-layer perceptron to predict the classes of success, failed, live and canceled based on that we can find the attribute used for the prediction of theses classes. We achieved 80 percent which is less as compared to the work done before where they achieved [23] 93 percent accuracy in the binary classification problem of success and failure contribution in the year 2018.

But this idea is new as we know that binary classification of success and failure contribution is old, but it now improved. Similarly, this idea is new but then people will come forward to achieve more accurate result.

We also used gradient boosting machine algorithm for the down sampled data which represent 10 percent of the data and achieved 87 percent accuracy in the test set. As we used the gbm function in R , it is too slow and more main memory and fast processor as well as GPU required to achieve goal or go for cloud. AS we have tested this and have a promising result. More work is required on this domain.

ACKNOWLEDGEMENT

We would like to thank our supervisor **Prof. Omair Shafiq** for their valuable guidance and motivation which was essential for the progress and completion of this term project report. Throughout the term your valuable guidance, tips and advises helped me in completing this task, and provide a clear and detailed picture of the topic is very helpful and guide us in our academic and professional life.

REFERENCES

[1] Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. Journal of business venturing, 29(1), 1-16.

[2] Kaggle (2018). Kickstarter projects [Data file]. Retrieved from https://www.kaggle.com/kemical/kickstarter -projects

[3] Belleflamme, T. Lambert and A. Schwienbacher, "Crowdfunding: Tapping the Right Crowd", SSRN Electronic Journal, 2012.

[4] A. Cosh, D. Cumming and A. Hughes, "Outside Enterpreneurial Capital", The Economic Journal, vol. 119, no. 540, pp. 1494-1533, 2009.

[5] A. Agrawal, C. Catalini and A. Goldfarb, "The Geography of Crowdfunding", SSRN Electronic Journal, 2010.

[6] V. Kuppuswamy and B. Bayus, "Crowdfunding Creative Ideas: The Dynamics of Project Backers in Kickstarter", SSRN Electronic Journal, 2013.

[7] G. Ahlers, D. Cumming, C. Guenther and D. Schweizer, "Signaling in Equity Crowdfunding", SSRN Electronic Journal, 2012.

[8] V. Nocke, M. Peitz and F. Rosar, "Advance-purchase discounts as a price discrimination device", Journal of Economic Theory, vol. 146, no. 1, pp. 141-162, 2011.

[9] Danhong Chen, Tianyu Yi," Establishment and Operation of College Students Entrepreneurship Crowdfunding Website Based on the Crowdfunding Mode", 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, 2016

[10] Lester Allan Lasrado, Artur Lugmayr, "Equity Crowdfunding- A Finnish Case Study", IEEE conference on Multimedia and Expo Workshops (ICMEW), 2014

[11] Liu Xuefeng, Wang Zhao," Using Crowdfunding in an Innovative Way: A Case Study from a Chinese Crowdfunding Platform", Portland International Conference on Management of Engineering and Technology , USA, IEEE, 2018

[12] Rafat M. Abushaban, "Crowdfunding as a Catapult for Innovation in the Middle East: Obstacles and Possibilities", IEEE Global Humanitarian Technology Conference (GHTC), pp. 433-440, 2014

[13] Meilin Shi, Lei Guan," An Empirical Study of Crowdfunding Campaigns: Evidence From Jing Dong Crowdfunding Platform", 13th International Conference on Service Systems and Service Management (ICSSSM), IEEE, pp.1-5, 2016

[14] Zhanyang Song, Danhong Chen, Tianyu Yi," The Research on Construction and Application of College Students Entrepreneurship Crowdfunding Platform", 2016 Chinese Control and Decision Conference (CCDC), IEEE, pp.4483-4486, 2016

[15] Yusan Lin, Wang-Chien Lee, "Analysis of Rewards on Reward-Based Crowdfunding Platforms", 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp.501-504, 2016

[16] Kevin Chen, Brock Jones, Isaac Kim, Brooklyn Schlamp, "KickPredict: Predicting Kickstarter Success", 2016

[17] Zhiyuan Tian, Lei Guan, Meilinshi, "The Key Factors of Successful Internet Crowdfunding Projects – An Empirical Study Based on Different Platforms", 2017

[18] Francisco Ferreira, Leandro Pereira," Success factors in a reward and equity based crowdfunding campaign", IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2018

[19] Peng Du1, Lei Xu1, Xiaojiao Qiao1, Haobo Li1, Dahui Li, "Post-crowdfunding in Reward-based Crowdfunding with Strategic Purchasing Consideration", International Conference on Service Systems and Service Management, 2017

[20] Jaya Gera, Harmeet Kaur," Influence of personality traits on campaign success", 8th International Conference on Cloud Computing, Data Science & Engineering , 2018

[21] Yani Li, Vineeth Rakesh, Chandan K. Reddy, "Project Success Prediction in Crowdfunding Environments", 9th ACM International Conference on Web Search and Data Mining, 2016

[22] Sharma, S. (2017). activation Functions: Neural Networks. Retrieved from https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[23]. Pi-Fen Yu, Fu-Ming Huang*, Chuan Yang, Yu-Hsin Liu, Zi-Yi Li, Cheng-Hung Tsai," Prediction of Crowdfunding Project Success with Deep Learning", IEEE 15th International Conference on e-Business Engineering (ICEBE), China, 2018

[24] Computational Complexity of Neural Network retrieved from https://kasperfred.com/posts/computational-complexity-of-neural-networks

[25] Time complexity question retrieved from https://ai.stackexchange.com/questions/5728/what-is-the-time-complexity-for-training-a-neural-network-using-back-propagation