EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Decision Tree Based Crowd Funding for Kickstarter Projects

Veena Grover[1,*], A. Anbarasi[2], Siddhesh Fuladi[3], M. K. Nallakaruppan[4,*]

[1]Department of Management Studies, Noida Institute of Engineering and Technology
[2]Department of Computing Technologies, SRM Institute of Science and Technology
[3]School of Computer Science and Engineering, Vellore Institute of Technology, Tamil Nadu
[4]School of Computer Science, Engineering and Information Systems, Vellore Institute of Technology, Tamil Nadu

Abstract

The proposed work employs the C4.5 decision tree algorithm on a kick-starter project dataset to help a user decide whether to back a kick-starter project that is ongoing by predicting how likely it is that it may be a successful one. We pre-processed the kick-starter dataset with about 35 columns, and used WEKA to run the algorithm on the dataset. We reached an accuracy of 99.7% and we also talk about why the algorithm chose 5 particular attributes over the others. A lot of other papers have discussed this problem from a project creator's standpoint, predicting whether a project is going to be a success before it has begun. There are fewer papers which look into predicting the success of the ongoing projects that helps users choose potentially successful projects to back, and we have also achieved a higher accuracy rate.

Keywords: Kickstarter, Decision trees, Crowdfunding

Received on 14 October 2023, accepted on 11 December 2023, published on 19 December 2023

doi: 10.4108/eetsis.4639

1. Introduction

Crowd funding is a practice of accumulating funds for a project or a venture by raising small amounts of money from a large amount of people in return for some form of incentive or reward. People, addressed as "backers", generally sign up to back these ventures in return for certain benefits like preferential treatment in sales or in some cases even large amounts of discounts. This also comes with the risk of the project that is backed failing to achieve its goal and being discontinued and the user wasting his money.

In order to develop an idea or a project, the developers need capital to fund them. Banks consider all SMB ventures and Entrepreneur start-ups very risky and don't provide loans against their venture because of economic downturn and today's restrictive lending policies. By this way there is very little business development, innovation or growth. A solution to this issue is crowdfunding. Kickstarter follows a pre-order model, which involves people making online pledges by investing in the product for development and get the product later for delivery. There are other models too, such as the reward-based model which is where the investor gets the satisfaction of helping and get reward. Naturally, the most important part of a crowdfunding project became attracting backers.

Now, in the point of view of a backer, one cannot decide if a project is worth backing by just what the product website shows them. Crowdfunding revenue projects invest in crowdfunding. While launching a campaign they set the funding goal and a deadline. The investors also get rewards from the user either as an acknowledgement from the backer or can also be deeply involved in the product's design. In fundraising the product is said to be successful if it only reaches the funding goal before the deadline or the campaign is considered a failed one and no money is involved. This paper proposes a system which will help potential backers to identify projects which are on their way to success, this

*Corresponding author. Email: nallakaruppan.mk@vit.ac.in

system works with incomplete data gathered from ongoing projects and attempts to predict the future state of the project. Many tools address this issue from the project creator's standpoint, but the backer's view is an important one to take into consideration too, all potential backers will want to reduce their risks by not backing projects with are on the path to failure, this system helps backers in making smart choices on the projects to back. According to the statistics less than 50 percent of the campaigns succeed. So, the campaign with the clear success rate spends more time on perfecting the product, extending the goals and the failing ones try to increase the reach of the product via all other social media so that it can be brought back to fruition. funding was projected to be over 2.8 billion dollars in 2012 and more than 16.6 billion dollars in 2014. The growth of crowdfunding is very similar to the growth of the internet in the early 90s. Figure 1. displays the flowchart for the approach used.

2. Dataset source

Kaggle is an online platform of data scientists and machine learners owned by google. It allows data scientists all over the world to explore and publish datasets. Data scraped and hosted on Kaggle is available free to use for anyone in the world. The dataset [1] was mined from the online website of Kickstarter [2] by the data scientists at the online community of Kaggle [3]. It contains comprehensive set of features and it spans across 18,142 records of projects from the years 2013 - 2015. The said dataset contains 35 features (columns). There are a wide variety of columns which span different dimensions of the data such as the financial side and the social side. In this section, we describe the various features that are present in the dataset through Table 1 and we explore the trends in the dataset using different visualization techniques

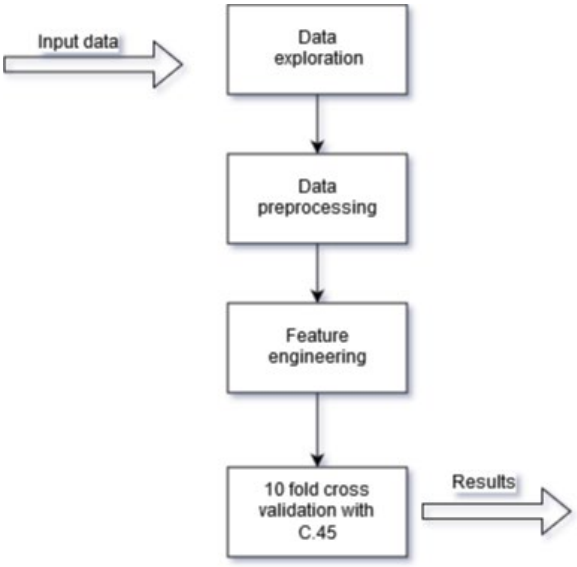

Figure 1. List of the Methods

Table 1. List of attributes, their description and datatypes

Attribute Name	Description	Datatype
State	This column is considered as the target class for the prediction algorithm, it consists of two attributes	String
Top Category	Since the service of kickstarter is available worldwide, the projects pitched use different monetary and fiscal mediums and this requires us to convert all currencies to a singular unit of exchange.	String
Updates	The column gives information on the number of times the details of regarding the project are updated by the creator.	Integer
Comments	This column provides number of comments on a particular project. This can provide information regarding to the popularity of the project in the social platforms and communities.	Integer
Rewards	This describes the number of reward tiers that are offered to potential backers, more reward tiers imply more incentive for potential backers. Usually, more reward tiers also imply a more diversified product range that might have a hand in the future state of a project.	Integer
Goal	This is one amongst the most prominent features of the dataset, it describes the the amount of funding that is requested by the creator of the project to successfully complete the project. This column is populated in the currency described in the currency tab.	Integer
Pledged	This column describes the amount of money that was amassed in total by the project for the duration of the time that it was active. This column is populated in the currency described in the currency column.	Integer
Backers	This column is also one of the fundamental features which help in determining the state of the project. It describes the number of backers (uses who enrol in a reward tier) that the project amassed for the duration of the time that it was active.	Integer
Duration Days	This column describes the number of days the project was active. This column in useful in describing the rate at which the project raises money.	Integer
Facebook Connected	This is a Boolean value which describes whether the project, that is being described, is connected to Facebook or not. This helps determine the social outreach of a particular project.	Boolean
Facebook friends	This column describes the amount of people that follow the project on Facebook. This column contains null values if a project has not been connected to Facebook.	Integer
Facebook shares	This column describes the Amount of people that have shared the project on their profile in Facebook. This describes deep personal interest of the people in the project. This column contains null values if a project has not been connected to Facebook.	Integer
Has video	This is a Boolean value which describes whether the creator has included a video in the description of the product. This usually helps in determining the legitimacy of a project.	Boolean
Creator projects created	This column dictates the number of projects that the creator has pitched apart from the project in consideration. This will help to determine the amount of experience that the creator has with respect to creating projects.	Integer
Videos	This column will contain the number of videos used in the description of the project. This column is filled with null values if the corresponding field Has Video is set to false.	Integer
Number of Images	This column contains the count of the images that are used by the creator in the description to capture the interest of potential backers in their product.	Integer
Number of words in description	This column describes the number of words used by the creator to describe their project aside from other media-based description	Integer
Number of words in risks and challenges	This column describes the number of words used by the creator to describe the risks and challenges associated with the project.	Integer
Number of FAQs	This column describes the number of words the creator used to address the questions asked by the potential backers regarding the project, this feature reveals a lot about the amount of interest shown by the users in the project.	Integer

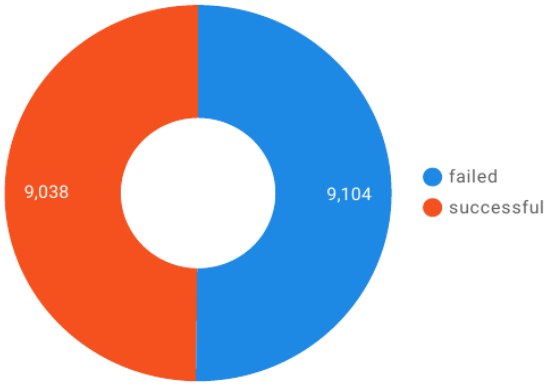

Figure 2. Total number of failed and successful projects.

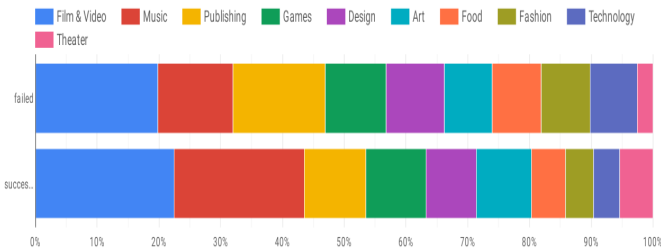

Figure 3. Failed and pass percentages in each top category.

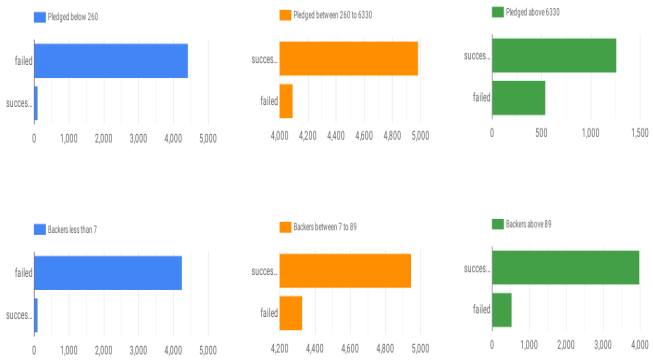

Figure 4. No of projects in each quartile of pledged and backers divided by state.

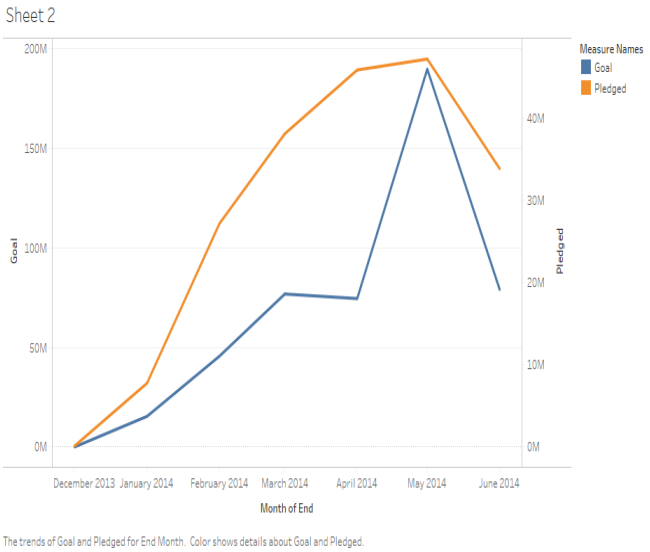

Figure 5. Goal vs. pledged for the duration of December 2013 to June 2014.

Figure 2 splits the projects on the basis of their state, this displays how skewed the data is to a particular attribute. Figure 3 depicts the ratios of the categories of projects in each state. Figure 4 classifies the number of successes and failures in each quartile this. Figure 5 plots the amount of money pledged and the goal over the time during which the projects were active.

3. Data Preprocessing

The raw dataset had many inconsistencies and missing values. We pre-processed the data by removing nulls, special characters. A small python script written using the pandas library, which was used to replace these values with zeros [4]. The nulls in Facebook Friends were replaced with zeros because if the Facebook Connected is a Boolean false, then we cannot replace the nulls with averages, because it would be an inconsistency.

Columns which do not contribute to the prediction of the final state, like website URLs of the project, were removed from the dataset. All the currencies in the dataset were in region specific so they had to be converted to a standard of currency and in our case, it was US Dollars.

Decision tree algorithm:

In this implementation, we use the C4.5 algorithm, which is an extension of the more known ID3 algorithm for decision trees. This was developed by Ross Quinlan and was first mentioned in his seminal Machine Learning journal article, published in 1994. The process of constructing a decision tree involves calculating the gain of splitting the data based on certain attributes. It calculates the information gain by finding the difference of the entropies of the parent and each potential child node. Essentially, if an event is very probable, it gives no new information to the tree. So, the entropy would be very

high and the information gain would be very low. Here, a very probable event refers to the situation of most of the records with a certain attribute value falling into one particular target class. We calculate entropy by using the formula in equation (1):

$$\boldsymbol{entropy}(\boldsymbol{T}) = \sum_{i=1}^{n} \boldsymbol{p}(\boldsymbol{x}) \log_2 \boldsymbol{p}(\boldsymbol{x}) \quad (1)$$

We can calculate information gain by using equation (2):

$$\boldsymbol{information\ gain}\ (\boldsymbol{T}, \boldsymbol{X}) = \boldsymbol{entropy}(\boldsymbol{T}) - \boldsymbol{entropy}(\boldsymbol{T}, \boldsymbol{X}) \quad (2)$$

C4.5 is called a statistical classifier since it assigns new incoming observations to mutually exclusive classes [5]. It differs from the ID3 algorithm as it uses single pass pruning i.e., removing features that are not useful for classification, to avoid overfitting. C4.5 handles issues that ID3 could not. C4.5 can classify both discrete and continuous data, handle missing attribute values by not using them in gain and entropy calculations and pruning trees after they have been generated, to delete any branches which are not useful in classification and replacing them with leaf nodes [6].

WEKA has implemented the C4.5 in java as J48, which is the algorithm we have used for our experiment, along with the 10-fold cross validation technique [7-8]. This evaluation technique involves dividing the dataset into 10 parts (or folds), training the model on 9 of them and testing on the remaining 1 [9-10]. After cross validation and computation of the evaluation results, the algorithm is invoked a final 11th time on the entire dataset and obtains the final model [11].

4. Results

The tree that resulted had 60 leaves and a size of 120 altogether. The features selected by the decision tree for prediction are Goal, Pledged, Backers, Number of FAQs, and Number of words in description. The information gain was manually calculated and is visualized in **Figure 6**.

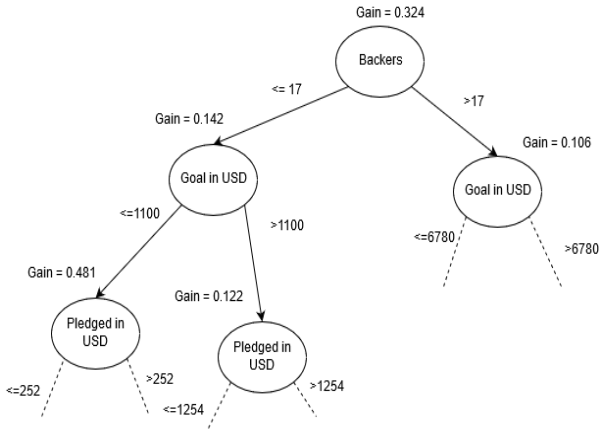

Figure 6. Top nodes with corresponding information gain values.

The resultant confusion matrix is shown in Table 2 and the classification parameters are shown in Table 3.

Table 2. Confusion matrix

	Predicted successful	Predicted failed
Actual successful	9030	74
Actual failed	22	9016

Table 3. Classification Parameters

Accuracy measure	Value
Kappa statistic	0.9894
Mean absolute error	0.0072
Root mean squared error	0.0715
Relative absolute error	1.4451 %
Root relative squared error	14.3024 %
Accuracy	99.4708 %
Precision	99.7569%
Recall	99.1871%
F1 score	0.9947

Possible reasons for the selected features:
The fields Goal, Pledged, and Backers are primary factors which affect how a project performs on a basic level. They decide the popularity and the amount of money amassed by the project, which is crucial in determining the state of the project.

The use of the field Number of FAQs to the prediction is two-fold, it shows the interest the community has in a particular project, it also shows the community engagement on the part of the creator. It informs a potential backer about the common facts about the project and this gives them more information to base their decisions. This is a possible reason why the algorithm chose this feature.

An increase in the value of field Number of words in Description will make the project simple and understandable to potential backers, and maybe that is the reason it attracted more backers and rendered the project successful.

5. Conclusion

This paper talks about crowdfunding, more specifically kickstarter, and analyses data from the website to generate a decision tree using the C4.5 decision tree algorithm. It talks about the possible reasons why the algorithm decided to choose the features it chose to classify the data. This paper proposes a system which implements a decision tree which uses the features Goal in USD, Amount pledged in USD, Number of Backers, Number of words is FAQ and Number of words in description. The system takes inputs of on-going projects with these features which help it to predict the future state of the project. An application of this would be a complete tool that potential backers could use to check which project they would like to back. There are some online tools such as Canhekickit and kicktraq that help project creators to tune their projects for it to be successful, but none of them have been proper success predictors.

References

[1] Patil, S., Mehta, J., Salunkhe, H., & Shah, H. (2021). Kickstarter Project Success Prediction and Classification Using Multi-layer Perceptron. https://doi.org/10.1007/978-981-33-4087-9_60

[2] Kuppuswamy, V., & Bayus, B. (2015). Crowdfunding Creative Ideas: The Dynamics of Project Backers in Kickstarter. SSRN. Retrieved from https://ssrn.com/abstract=2234765

[3] Hornuf, L., & Cumming, D. (Eds.). (2017). The Economics of Crowdfunding: Startups, Portals, and Investor Behavior. Forthcoming. SSRN. Retrieved from https://ssrn.com/abstract=2234765 or http://dx.doi.org/10.2139/ssrn.2234765

[4] Agyeah, G., Mark, B., Adesiyan, J., & Kolomoytseva, A. (2019). Modeling the Success of Kickstarter Projects.

[5] Kaggle. (2019). Kickstarter Dataset. Retrieved from https://www.kaggle.com/tayoaki/kickstarter-dataset

[6] Kickstarter. (2019). Retrieved from https://www.kickstarter.com

[7] Kaggle. (2019). Retrieved from https://www.kaggle.com/datasets

[8] Quinlan, J. C. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.

[9] Wikipedia. (2019). Statistical Classification. Retrieved from https://en.wikipedia.org/wiki/Statistical_classification

[10] Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A Comparative Study of Decision Tree ID3 and C4.5. International Journal of Advanced Computer Science and Applications, 4(2).

[11] Frank, E., Hall, M. A., & Holmes, G. (2016). The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.