Big Data Continual Assessment

Kickstarter projects success predictions

Antti West

Institute of Technology Tralee, Dromtacker, Tralee, Co. Kerry
Ireland

Abstract

Kickstarter is the most popular crowdfunding platform on the internet. It offers people and companies, big or small alike, a chance to gather funds for their projects through crowdfunding. This report's main goal is to find out what kind of crowdfunding projects are more likely to succeed with their funding. There are many other crowdfunding platforms out there but because Kickstarter is the biggest one, this report only focuses on the data of projects from Kickstarter.

Keywords: Kickstarter, crowdfunding, success, predictions

1. Introduction

New projects and ventures need resources such as financing and publicity to succeed. That's where Kickstarter comes in. Kickstarter is the world's largest funding platform for all kinds of creative projects. Kickstarter projects are commonly put up by single person or a small group of creative people, but sometimes bigger companies search funding through crowdfunding instead of seeking out venture capital e.g., for a specific project or a product which they are aiming to bring to life. The magnitude of the funded projects may vary from a small idea of a one person with funding goal of a few hundred dollars to a huge venture of a big company searching for an investment with a funding goal of hundreds of thousands of dollars. The most common way of seeking crowdfunding is to put up an introductory demo or a preview of a project that is being funded and market it as attractively as possible and to make it appealing to a mass of people to back the project.

The Kickstarter platform has over 140 000 funded projects to this day and has over 14.5 million backers in total for all the funded projects (Kickstarter, 2018) since its establishment in April 2009. Kickstarter has kickstarted many projects with huge amount of funds donated to creative ventures such as a 3D printer designed to be affordable for the average consumer called The Micro, which has raised over 3,4 million dollars and a smartwatch from the company Pebble Technology called Pebble Time which raised over 20 million and is the most funded project in Kickstarter (Kickstarter, 2018). With all the successful projects in the platform there is even more unsuccessful projects that for one reason or another have failed before they were unable to reach their funding goal. Alongside failed and successful projects there are cancelled projects and projects who never delivered their promise or products even though they reached their funding goal. These projects decrease the reliability and credibility of crowdfunding and the Kickstarter platform but are an inevitable by-product of the loosely supervised funding system.

Even though crowdfunding is a working way to get investments to a project and around it revolves millions of dollars, very little academic knowledge has been recorded or gathered about it (Agrawal, Catalini, & Goldfarb, 2010). In M. Rice's *Co-production of business assistance in business incubators: an exploratory study* is an effort of collecting analytical understanding of the nature of crowdfunding. This analysis suggests that usually the projects crowdfunding goal is passed by a narrow margin, but the failure is usually by a huge amount (Rice, 2002). The study also suggests that delays are a common thing to happen in a crowdfunded project and this can be predicted from the size of the project (the bigger the project more likely it is to be delayed). All these factors and many more add to the challenge for person or company who is trying to get their project crowdfunded.
Crowdfunding isn't easy and by no means is a sure way to get projects funded or collect money. That's why this report's aim is to find out through the tools of data-analysis what kind of projects are more likely to succeed and

what kind of attributes successful projects usually have in common and what similarities failed projects have and what separates these two. The dataset only contains nominal, binary, ordinal and numerical data so evaluating the quality of the preview video, the sales pitch or any other means of marketing besides the information on the dataset is discarded in this report.

2. Methods

The analysis was started by importing the Kickstarter dataset from Kaggle.com and then read with Pandas framework. The KDD process was started by studying the dataset. The length of the newly imported data was 378661 rows and it had 15 columns: ID, name, category, main_category, currency, deadline, goal, launched, pledged, state, backers, country, usd_pledged, usd_pledged_real and usd_goal_real. The category and main_category columns define the type of the project such as main_category is publishing, and category is poetry e.g. The currency column tells the currency of the donations and it also can indicate from which country the project is based on. Deadline and launched columns are self-explanatory. The goal and usd_goal_real columns define the amount of funding or donations the projects are aiming for. The pledged, usd_pledged and usd_pledged_real columns show the amount of gathered funds for the project. On the columns that define the wanted goal and the amount of pledged funds we use only the usd_pledged_real and usd_goal_real since they are converted to USD and can be equally compared. Each state column has a status of either successful, failed, canceled, live or suspended. The value of the state column depends on if the project had reached their goal due to the deadline so if the usd_pledged_real exceeds the amount of usd_goal_real at the date of deadline the status is successful. But in some cases, the status has set to canceled although the goal has been reached. This must be taken in to notice during the data-analysis. The backers column tells the number of people who had funded the project and the country column shows that which country is the project from.

After looking through the dataset the next step was to start preprocessing the data. The data had some null values on it and some unnecessary columns for the reports purpose such as usd_pledged, pledged and backers. Backers is not needed because it might cause problems with the datamining process later. The only null values after dropping the usd_pledged column was on the 4 values on the name column, but since the name is also irrelevant for this study they were kept in the dataset as empty values marked as NaN. The reports main objective is to find out if the project is going to be successful with its funding or not so the state columns ordinal values were changed to plain boolean values such as 1 or 0, 1 indicates that the project's funding was successful and 0 meaning that it failed and drop the rows which had either live or suspended value on their status column. This caused the rows of data to drop from 378661 rows to 331675 rows which is roughly a 13% drop from the original amount. Due to the fact that some projects were cancelled and still met their goal another column had to be created called goal_met. This column indicates the true state of the funding. The date columns were string values so for easier handling of dates the launched and deadline columns were converted to datetime format. The dataset also included some records where the date was clearly wrong such as date se to 1970's etc. so to further validate the credibility of the data those rows were dropped as well.

The next part of the KDD process was to gather basic information about the dataset. This was done by calculating averages and mean values across the dataset such as average success rate of projects in each category and country. Also, information about the timeframes of which the projects were funded, the number of projects emerging from each country and how much funds in total were donated to the projects by category and country were gathered.

The next step focused on the timeframes between the launch date and the deadline of the project. The mean of these values was calculated by splitting the data into failed and successful projects and the calculating the mean of the difference in days between deadline and launch date of the datasets. After calculating the time differences, the attention was shifted to countries. By far the country with the most projects, successful or failed, was United States with 78% of the whole number of projects started. The two next biggest countries were Great Britain at 9% and Canada with 4%.

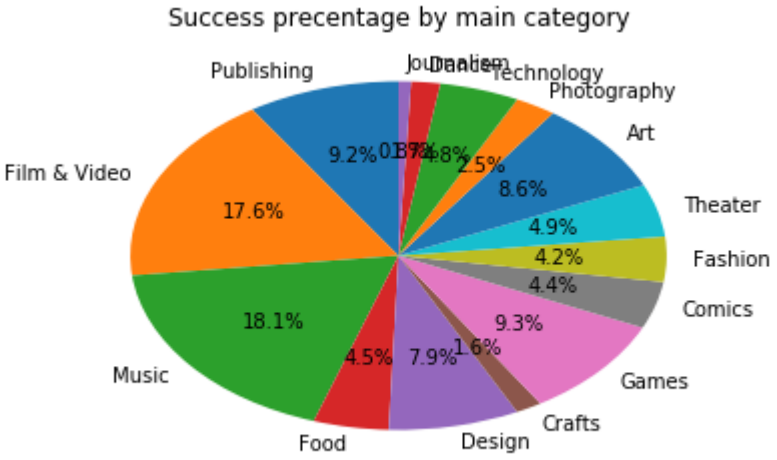
After studying and visualizing the data one observation was that category made a huge difference on the project's success. Taking this into account the data was trimmed so that all the main categories had their own column so that it could be fed through a decision tree algorithm. For classification purposes the predictors for the algorithm were the main categories and target was the goal_met column. After defining the predictors and the target that data was split into training and testing sets. The division was made with 70/30 ratio where 70% of the data was set to training data and the rest 30% was inserted into the test portion. The data was divided randomly by creating as many uniformly distributed random numbers between 0 and 1 for each row of data in the dataset. This way the training and testing data is more reliable since its order is randomized. This is because in some cases the values of rows close to each other might have an influence on each other so randomizing the split decreases the risk of unreliability. After splitting the data, the algorithm was "trained" using the training set.

After first running the decision tree with just categories and because of its poor accuracy results, the next step was to give more detailed data to the algorithm. Timeframe of the funding and goal to goal mean ratio columns were created in effort to summarize the data more precisely. These two columns and the country column were added to the next iteration of the algorithm.

The decision tree was then visualized by creating a .dot file of the logical process of the algorithms reasoning of the data and converting that to an image file. Due to multiple different variables the results were based on, the visualization image of the decision tree was huge, so it wasn't added to this report.

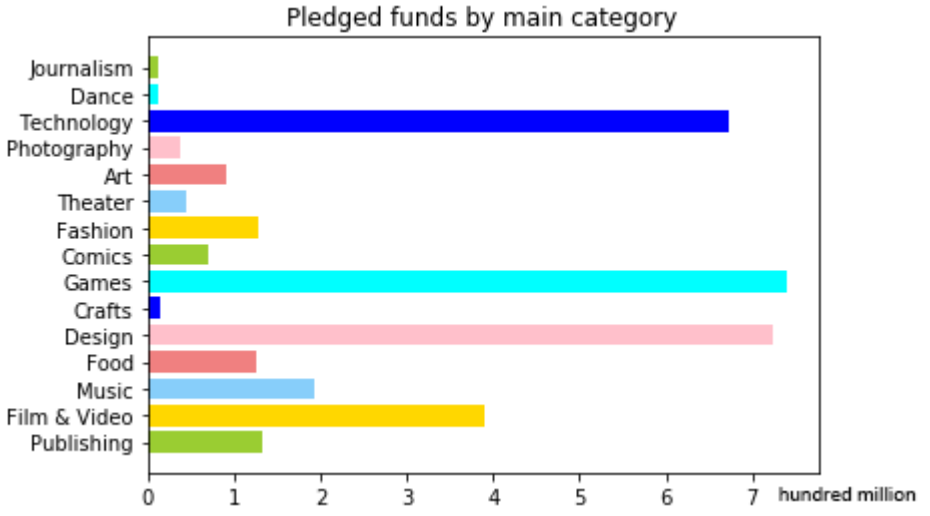
3. Results

By category the most successful projects where musical projects, however this category wasn't the most funded one. The most funded category of projects was Games. The average success rate of the was 36.35% so the majority of the projects fail to meet their funding goal. The mean values for successful projects funding time was 31.2 days and for the failed projects the time was 34.7 days. By main category the most successful one was music with 18.1% success rate and the rest of the top five most successful categories were Film & Video (17.6%), Games (9.3%), Publishing (9.2%) and Art (8.6%). The main category which had the most failed projects was Film & Video with 29% of failure rate and the two second most failed categories were Publishing (20%) and Technology (19%).



The most pledged projects by the main category were Games with 739 million USD, Design (724 million USD), Technology (672 million USD), Film & Video (389 million USD) and Music with 193 million USD. From bottom to top the categories which received the least amount of donations in total were Journalism with 12.3 million USD,

Dance (12.9 million USD), Crafts (14.3 million USD), Photography (38.2 million USD) and Theater with 43.5 million USD.



The accuracy results from the first iteration of the decision tree algorithm, in which the only predictive factor was category, were poor. After running the algorithm for the second time and with more accurate predictions, the results were slightly better:

Only category data give to the algorithm:

Predictions	0	1
Actual		
0	68834	1897
1	37871	2477

Country, timeframe and goal to goal mean ratio added:

Predictions	0	1
Actual		
0	56408	14143
1	24499	15815

In the tables above the number 0 means failed project and 1 means successful. The first table can be interpreted to read so that predictions for failed projects were right in 68834 times out of 70731 of total count of failed projects. This gives us an accuracy rate of 97.3%. On the other hand, only 2477 of the total of 40348 of the successful projects were predicted correctly with an accuracy rate of only 6.1%. On the second table the numbers are more balanced. The prediction accuracy for failed projects dropped from 97.3% to 79.9%, but the accuracy for successful projects rose from 6.1% to 60.7%.

4. Conclusion

Even though the results from the decision tree got slightly more accurate when adding more predictive elements to the mix, the results were still not nearly accurate enough to assuredly say that when given a projects category, country of origin, amount of time for collecting the funds and the projects goal-to-mean-goal ratio if the project will or will not succeed in its funding. One reason for this is the fact that the data didn't contain information about the projects marketing, presentation or overall attractiveness, which are possibly even bigger factors to the project's success then the ones mentioned before. However, based on these results, the most funded projects are not necessarily the most

successful ones e.g. the most successful category of projects was music projects but its 5th on the most funded categories.

References

Agrawal, A., Catalini, C., & Goldfarb, A. (2010). The Geography of Crowdfunding. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1692661

Rice, M. P. (2002). Co-production of business assistance in business incubators: An exploratory study. *Journal of Business Venturing*, *17*(2), 163–187. https://doi.org/10.1016/S0883-9026(00)00055-0

Data of Kickstarter's total of funded projects, backers and most funded projects based on information on www.kickstarter.com.