Capstone Final Report

Kickstarter Crowdfunding Success

October 2017

Problem

[1]One of the biggest challenge for new entrepreneurs according to Entrepreneur magazine, is funding a new business. As such, entrepreneurs look for all possible funding solutions to finance their endeavors.

There are several ways to deal with it from funding everything by one's self, getting loans or finding investors each having there pros and cons.

Another option available is crowdfunding. [2]"In 2015, it was estimated that $34 billion was raised in this way". [3]"Crowdfunding is the practice of funding a project or venture by raising monetary contributions from a large number of people".

Kickstarter is one of these corporations that maintain a global crowdfunding platform having received more than [4]$3.3 billion dollars pledged to its projects. Its mission is to [5]"help bring creative projects to life". It uses an "All-or-Nothing" approach, meaning projects only get the pledged amounts when the funding goals are reached.

Of course, not all projects will be successfully funded in the platform. [6]According to Kickstarter itself, success rate with the platform ever since it started is just a little bit more than 35%. The goal of this study is to find ways to increase the probability of getting funded.

Client

This study is done for the entrepreneurs and individuals who have been turning to Kickstarter as an option to fund their endeavors. It seeks to increase their chances of getting funded successfully. As a result, these groups of individuals will make modifications in their projects to improve their chances of getting successfully funded.

---

[1] https://www.entrepreneur.com/article/254721

[2]https://www.forbes.com/sites/chancebarnett/2015/06/09/trends-show-crowdfunding-to-surpass-vc-in-2016/#34f98c7e4547

[3] https://en.wikipedia.org/wiki/Crowdfunding

[4] https://www.kickstarter.com/help/stats?ref=about_subnav

[5] https://www.kickstarter.com/about

[6] https://www.kickstarter.com/help/stats

Data Description

The data comes from kaggle, [7]"a platform for predictive modeling and analytics competitions in which companies and researchers post data and statisticians and data miners compete to produce the best models for predicting and describing data".

Specifically, it comes from the following link:
https://www.kaggle.com/socathie/kickstarter-project-statistics/data

It contains the top 4000 most backed projects on Kickstarter.

It is available in CSV format and includes the following information:

- amt.pledged
- blurb
- by
- category
- currency
- goal
- location
- num.backers
- num.backers.tiers
- pledge.tier
- title
- url

Here's an example of what the data looks like:

	amt.pledged	blurb	by	category
0	8782571.0	\nThis is a card game for people who are into kittens and explosions and laser beams and sometimes goats. \n	Elan Lee	Tabletop Games

Here's the continuation of the same data row:

	currency	goal	location	num.backers	num.backers.tier	pledge.tier
0	usd	10000.0	Los Angeles, CA	219382	[15505, 202934, 200, 5]	[20.0, 35.0, 100.0, 500.0]

And here is the last columns in the same data tow:

	title	url
0	Exploding Kittens	/projects/elanlee/exploding-kittens

[7] https://en.wikipedia.org/wiki/Kaggle

Another dataset that was used in the study, is the live dataset from the same source. It contains the top 4000 most backed projects on Kickstarter.

It is available in CSV format and includes the following information:

- amt.pledged
- blurb
- by
- country
- currency
- end_time
- location
- percentage_funded
- state
- title
- type
- url

Here's an example of what the data looks like:

	amt.pledged	blurb	by	country
0	15823	\nCatalysts, Explorers & Secret Keepers: Women of Science Fiction' is a take-home exhibit & anthology by the Museum of Science Fiction.\n	Museum of Science Fiction	US

Here's the continuation of the same data row:

	currency	end.time	location	percentage.funded	state
0	usd	2016-11-01T23:59:00 -4:00	Washington, DC	186	DC

And here is the last columns in the same data tow:

	title	type	url
0	Catalysts, Explorers & Secret Keepers: Women of Science Fiction	Town	/projects/1608905146/catalysts-explorers-and-secret-keepers-women-of-sf?ref=discovery

Data Wrangling

Here are the data wrangling steps done on the project data set to solve the problem:

A. *most_backed.csv* was read to data frame. This data is available on the /data folder. This dataset was mainly used for exploratory data analysis to answer the problem. Here are the steps taken to clean the data:
   1. An unnecessary column was deleted (*'Unnamed: 0'*)
   2. *'.'* where replaced using *str.replace* with *'_'*
   3. *'by'* column name was renamed to 'creator'
   4. All *'\n'* was removed from the blurb
   5. Removed all non-*usd* currencies to standardized the *amt_pledged* and *goal* figures
   6. *.groupby()* and *count()/sum()* was used to summarize the data columns for analysis
   7. For the *goa*l column, goals were rounded up to the nearest 5000 to create better groups

B. l*ive.csv* is the other dataset used for this study. This data is also available on the /data folder. This data set is used to create a machine learning model to predict Kickstarter results. Here are the steps taken to clean the data:
   1. Steps 1, 2, 4, and 5 from A was also done on the data.
   2. An *end_time* limit of 2016-12-01 was used based on the suggested [8]30 day campaign period by Kickstarter to filter the data based on the earliest end_time of 2016-10-29
   3. Created an *'s'* column to generate results
   4. Created a new data frame using the live data for machine learning purposes having *'blurb'* and *'s'* columns

C. [9]*TextBlob* was used to generate sentiment scores. [10]TextBlob sits on the shoulders of NLTK and another package called Pattern. It is a simple to use and boast a surprising amount of functionality. It produces polarity which outputs -1 to 1 with 1 meaning positive and -1 negative. It also produces subjectivity with 0 to 1 outputs, 0 being objective and 1 being very subjective. This output along with the *'s'* column was used for machine learning.

[8] https://www.kickstarter.com/blog/shortening-the-maximum-project-length

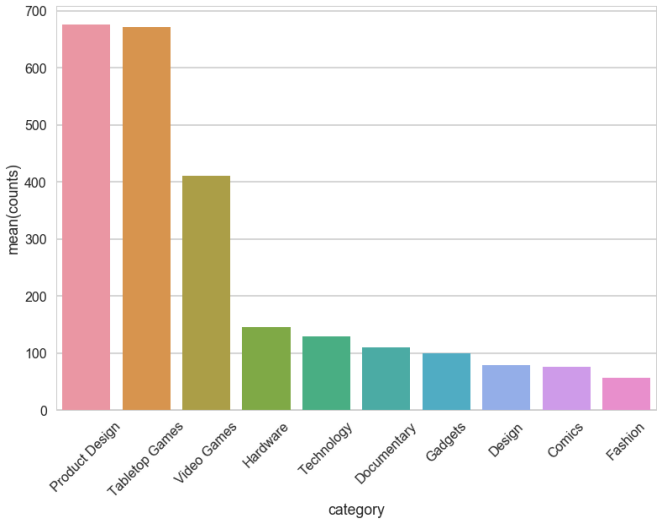
[9] https://textblob.readthedocs.io/en/dev/

[10]https://elitedatascience.com/python-nlp-libraries

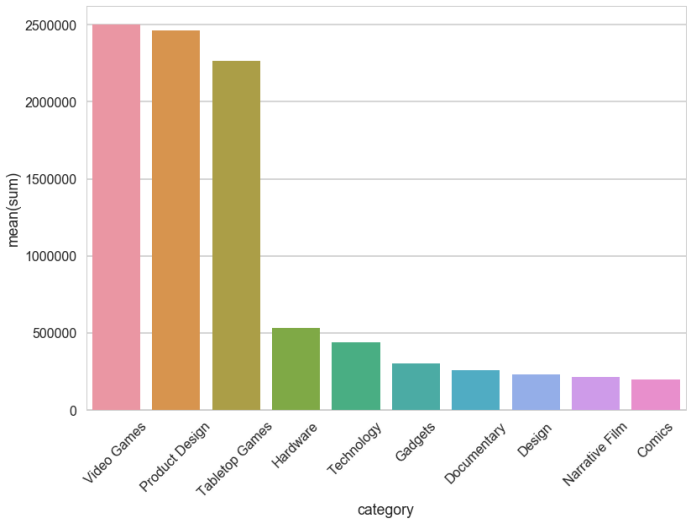
Data Statistics and Findings

Here are some Interesting statistics from the data:

Top 10 Most Backed Projects
by number of projects



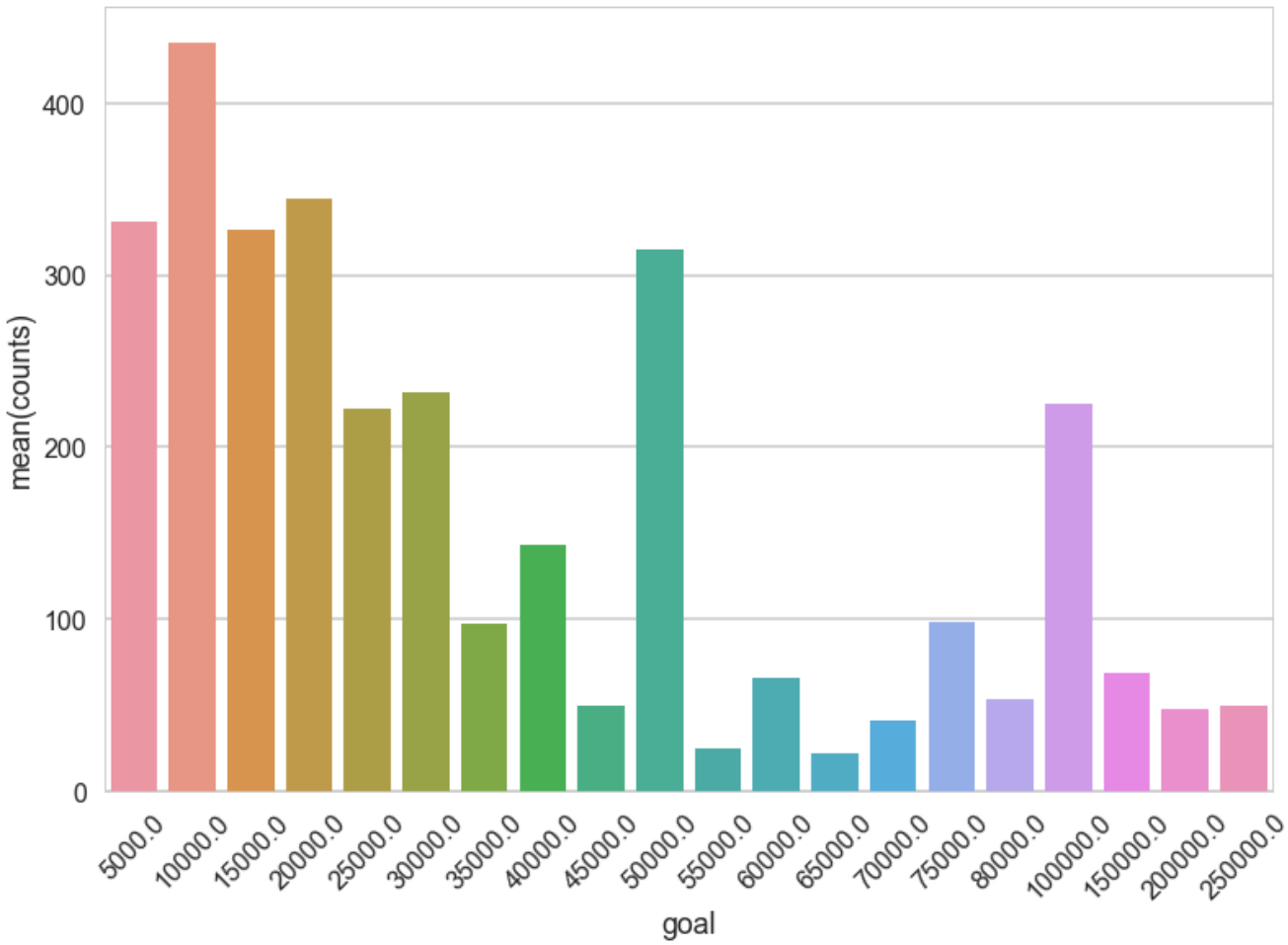
Top 10 Most Backed Projects
by number of backers



Though there were 114 categories in the data, the top 10 most backed projects by number of projects represents close to three-fourths of the most backed campaigns.

Furthermore, the most backed according to number of projects and number of backers are mainly 3 product categories and these are Product Design, Tabletop Games and Video Games.

This information provides us a a good idea of the profile of the backers in the platform and it can be used in formulating ideas to target market segments.



There were 310 different goals in the data. But after rounding up all the data to the nearest 5,000, it was narrowed down to 78 groups.

The table above shows the top 20 most backed goals. It represents more than 90% of the goals.

The top 5 goals on the other hand represent just a little over 50% of the projects. It has a wighted average of 19,000 which is pulled up by the 50,000 goal.

The smallest goal is 1 and the highest is 2,000,000. There were 4 projects with a 1 usd goal.

For all the data analysis and visualization, please refer to this link: https://github.com/raffyenriquez/Springboard/blob/master/Capstone%20Project/Capstone%20Project%202/Capstone_Final_Code.ipynb

Solution

In order to solve the problem, a lot of exploratory data analysis needs to be done on the most backed dataset.

Several features were looked at including the category, goals, locations and creators. These features were compared, counted, and summarized. Analysis was based on the assumption that it is representative of the entire population of Kickstarter campaigns. The author recognizes that sampling bias may play a factor on the data and can skew results.

A word cloud was also used to determine the common words in the title as well as for the blurbs. This information gave the author an idea of using text as a predictor of results.

For building a machine learning model, text sentiment analysis was used on the blurb. This was done by suing TextBlob. The polarity and subjectivity scores it provided was used as features to determined success.

The data that was used for this model was the live dataset. Another set of assumptions were made to use the data. *end_time* was used to determine a set campaign period and the limit was based on Kickstarter's [11]30 day campaign period recommendation and a study that said [12]80% reach Kickstarter goals in 3 days or less. The remaining data starts on October 29, 29016 and ends on November 30, 2016.

From here, success was determined by an article from Forbes stating [13]97% of campaigns fail to raise less than half of the funding. A qualifier of 50% funded was used to determine if a campaign was successful or not.

---

[11] https://www.kickstarter.com/blog/shortening-the-maximum-project-length

[12] http://www.rudebaguette.com/2014/09/24/10-kickstarter-figures-you-need-to-know/

[13] https://www.forbes.com/sites/suwcharmananderson/2012/07/17/secrets-of-success-hidden-in-kickstarters-numbers/#62c71d266713

This data was then split to training and testing sets that were used for various machine learning models.

The classifiers used were Logistic Regression, KNearest Neighbor, Decision Trees, Random Forest and Support Vector Classifier.

The models were scored on the training and testing data to determine determine accuracy. Precision, recall and f1 scores were also checked to see the models exactness and completeness.

Limitations

There are inherent limitations to this data:

1. The data was limited in terms of dates and other information, assumptions needed to be made to be as close to actual results as much as possible
2. The data can be improved by having more features that can be used to correlate to results. No additional web scraping was done to get these other features.
3. This study was limited to a week's work timeline as it is the author's goal to simulate urgency in times when a Data Scientist has to make do with all the resources available and produce useful results.

Results

Based on the author's analysis of the data, goals and target market play a role in a creator's success. Being strategic on these things can pay huge dividends for a project's success.

It must be emphasized that these results might have some sampling bias from the dataset as it only contained the most backed projects. It is possible that there were a lot of projects coming from product design and games that really did well compared to the others while other categories did just enough. It is also possible that there simply are a lot more projects on the product design and games categories than say theatre thus resulting in more per category count.

This same reasoning also applies to the goals as there might be a lot of creators that had lowers goals but less success than say fewer creators with higher goals and more success.

By creating a machine learning model based on sentiment analysis, we can say that it can potentially contribute to a creator's project success.

Here are the scores from the Machine Learning Models created:

	train	test	precision	recall	f1
Logistic Regression	0.6419491	0.63559322	0	0	0
KNearest Neighbors	0.7251059	0.62923728	0.4594594	0.0988372	0.1626794
Decision Tree	0.8532838	0.61652542	0.4579439	0.2848837	0.3512544
Random Forest	0.8268008	0.62923728	0.4819277	0.2325581	0.3137254
SVC	0.6419491	0.63559322	0	0	0

Though as can be seen from the results the models didn't perform quite as good as had been hope for. The best performing model was the decision tree and the random forest classifiers if we look at precision and recall. Ironically, these classifiers also overfit the training data.

Recommendations

There are several factors that can increase one's chance of improving success in raising funds through Kickstarter. The obvious ones based on the data are in terms of goal-setting and being strategic with the projects in identifying a target market.

Having said that, here are the author's recommendations to improve success not just in Kickstarter but also with crowdfunding in general:

- Set two goals before launching a campaign. The first one, a real goal, to be the campaign's main objective in terms of total revenue and the second one, to be the Kickstarter goal, an optimal goal where the number makes sense considering cost and one's willingness to work for it.
- Before starting a campaign, the creators need to ask themselves who will there clients be and from there look out crowdfunding platforms and see where most of their target is.
- Though not covered directly by the data, crowdfunding is a social platform that relies on a lot of other people to create success. With the advent of social media, an understanding of social psychology can be used to create a bandwagon effect for certain campaigns. This can also lead people who are on the fence to support a campaign just by the sheer number of other people doing it. This also highlights the importance of having momentum in the campaign and even using sentiment to spur it.

Suggestions for Further Research

There are quite a number of ways to improve the results of this study.

1. Turn the blurb into a word vector and see how that contributes to the result. Use that data paired with sentiment analysis to see results.
2. Gather more text related data in how creators pitch their campaign such as video messages, and product descriptions.
3. Use sentiment analysis on the reception of campaigns on social media.
4. Include other features such as time of launch, and goals.
5. Use more data that is available for scraping on the web.
6. Use one hot encoding for some variables such as category, location or days.