# Determinants of a Successful Crowdfunding Campaign

By Stefan Diener

*Abstract* — **Crowdfunding platforms have seen great success in bringing together fundraisers and private investors. However, despite the increasing popularity of these websites, only one in five crowdfunding projects reaches their funding goal. This paper investigates close to 70’000 US crowdfunding campaigns that were scraped from the crowdfunding platform Indiegogo, and examines three factors that are potentially related to funding success, namely, campaign category, geographic location, and the main campaign description or so-called "story". Regarding the first factor, I find that the market size and average success rates are strongly dependent on the campaign’s category. With respect to the second and third factor, a random forest model was trained to classify successful and failed campaigns. Here, analyzing the feature importance of the model indicated that both the campaign story as well as the geographic location of the campaign founders appear to contain information that predicts funding success. However, deeper causal analysis is left to future works.**

***Keywords—crowdfunding, success prediction, machine learning***

## I. Introduction

Over the past decade, crowdfunding platforms like Indiegogo have emerged as a prominent way for entrepreneurs and artists across the globe to raise capital online [1]. Every campaign specifies a funding goal and creates a campaign story, in which the campaign owner tries to convince potential backers to fund the project through a combination of text, images, and videos. From this common starting point, some campaigns achieve tremendous success. The currently most successful start-up "Babymaker" on Indiegogo was able to raise over 10 million GBP for its e-Bike project, in spite of asking for only 20’000 GBP. However, despite this huge potential, most campaigns never reach their funding goal and are instead required to refund their backers. Due to this all-or-nothing approach, one of the biggest challenges in the crowdfunding literature is finding out, which key factors lead to a successful campaign. [1][2][3]

In this paper, I investigate three factors that potentially contribute to a successful campaign in the US crowdfunding market on Indiegogo. The goal of this analysis is to derive insight, through exploratory data analysis and machine learning, which can help prospective entrepreneurs develop an optimal strategy for their crowdfunding campaign.

## II. Analytical Questions and Data

### A. Analytical Questions

The scope of this analysis covers three main areas summarized by the following questions:

- How do market size and funding success differ across campaign categories?
- Does geographic location play a role in funding success?
- Are there patterns in campaign title, description, or story that predict a successful campaign?

The first question examines the role of Indiegogo categories, as they are an intuitive way to segment the crowdfunding market. They classify all campaigns into one of 28 categories, where each category belongs to one of the overarching areas "Tech & Innovation," "Creative Works," or "Community Projects."

The second question aims to investigate the role of geographic location on funding success. Specifically, I analyze the influence of potential network effects by having other campaigns in the same area, as well as the effect of the local economy and size of the start-up sector.

The last question investigates the potential of text analysis to derive features from the predominantly unstructured text data that indicate a successful campaign.

### B. Data

The data used to answer these analytical questions was scraped directly from the Indiegogo website. From the 11th to the 17th of November 2021 the scraper continuously downloaded every campaign that was publicly featured on the website at this point in time.

Additionally, to assess the effect of the broader state of the US economy and the start-up sector in each US state, data from the U.S. Bureau of Economic Analysis on state GDP and venture capital investments is [4].

## III. Data (Materials)

The unprocessed data consists of two data frames that are merged in the subsequent data preparation steps. The first consists of scraped Indiegogo campaigns and containing 118’605 records of 11 features for each campaign. These include eight features of type string, and two numerical features, as shown in the table below.

*Table 1: Campaign Data Feature Overview*

| Type | Feature |
|---|---|
| str | Campaign title |
| str | Campaign description |
| str | Campaign story |
| str | Campaign location (e.g. "Brooklyn, NY") |
| str | Campaign category (28 in total) |
| str | Current funding amount, currency, and percentage of goal reached (e.g. "$612 USD raised 12%") |
| str | Nr. of campaigns previously launched the founder (e.g. "2 Campaigns)" |
| str | Time outstanding until the funding window closes (e.g. "2 days left" or "Ended") |
| num. | Number of videos embedded in story |
| num. | Number of images embedded in story |

The second data frame contains yearly information on GDP and dispersed venture capital for each US state from 1995 to 2019.

*Table 2: Economic Data Feature Overview*

| **Type** | **Feature** |
|---|---|
| num. | GDP per state |
| num. | Venture capital investments per state |
| num. | Share of VC investments of GDP per state. |

The data is characterized by an imbalance in the target variable, as only about one in five campaigns are successful in reaching their funding goal, as shown in figure 1 below.

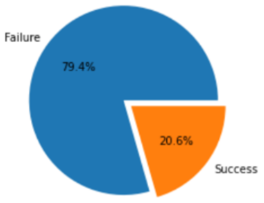

*Figure 1: Distribution Funding Success*

Additionally, the data is also coined by geographic imbalance. Although it includes campaigns from almost every country in the world, around 60% of all campaigns are based in the USA, whereas most other countries are represented by only very few campaigns, as shown below in figure 2.

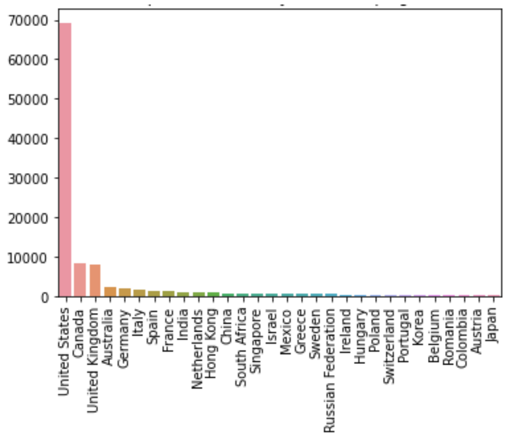

*Figure 2: Top 30 Countries by Nr of Campaigns*

Finally, a key characteristic of the data is that it consists in large parts of unstructured text data. Thus, the next section will explain the necessary steps that were taken to derive the final features that help answer the research questions.

## IV. Analysis

### A. Data preperation and derivation

The data preparation and derivation step can be separated into three main tasks. First is the basic cleaning and numerical derivations of the columns, second, the geographic preprocessing, and third, the text preprocessing.

#### 1) Basic cleaning and numerical derivations

The basic cleaning starts by converting the number of previous campaigns by the founder to a numeric value. Next, all campaigns that are unfinished, e.g., because they were still ongoing or did not launch yet, are dropped from the data. This ensures comparability between the funding success of different campaigns. Additionally, campaigns that had missing values in the campaign title or description were dropped, because those constituted cases where the campaign creator set up an incomplete campaign. This also ensures comparability between campaigns and facilitates the derivation of text features in later steps. In contrast, campaigns with missing values in the story were not erroneous but instead represent a story that only includes images. Hence, they were kept in the data and imputed by an empty string. Furthermore, the text-based funding column was used to extract the funding amount, currency, and percentage of funding goal reached. All funding amounts were converted to dollar equivalents, using exchange rates from yahoo finance. From this, the original funding goal and a binary success variable were derived. Lastly, the campaign category was used to infer the corresponding overarching category, i.e. "Tech & Innovation," "Creative Works," or "Community Projects".

The numeric funding amount, the funding success variable, and the campaign categories provide the quantitative basis for answering the first research question, as market size and success likelihood, can now be compared across categories.

#### 2) Geographic preprocessing

The processing of the address column required a multi-step approach. First, using the common formatting or the location string, the city and county component of the address was extracted by splitting the column at the comma. Next, a list of all U.S. states and their abbreviations was used to change the country of all campaigns that listed a state instead of the U.S. as their country to "United States." Subsequently, the data was filtered to only keep records from the U.S., reducing the size of the data by about 40%, leaving 69’389 observations in the data. Additionally, all addresses were geocoded using the Google Maps API. To minimize the number of costly API calls and to ensure reproducibility, a dictionary was created that associates each unique address with a geocoded object. This reduced the necessary API calls from 69’389 to only 7’891. This dictionary was used to extract the coordinates and corresponding state for each address in the data. With these coordinates Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is used to segment the data into multiple high-density clusters and one low density cluster. From this a new "urban" variable is constructed where each point that belongs to a high-density cluster has the value 1 and has 0 otherwise. The figure below illustrates this.

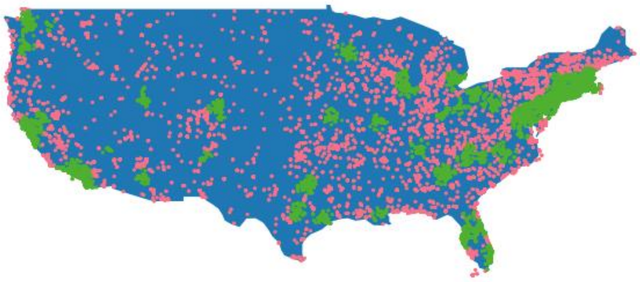

*Figure 3: High vs Low Density of Campaigns*

Furthermore, as now each campaign has a corresponding state, the GDP and venture capital data form the U.S. Bureau of Economic Analysis can be joined to the campaign data.

By deriving high-density clusters and joining data on GDP and VC investments per state, a robust data basis is created that allows to investigate the role of geography in the funding success, as suggested by the second research question.

### 3) *Text preprocessing*

Finally, the campaign title, description and story column are preprocessed in the same way. It starts with basic text cleaning, i.e., removing stop words, special characters, numbers, and converting all words to lower case. Then the TextBlob library is used to calculate a polarity and subjectivity score. The former measures the sentiment of a text, whereas the latter one quantifies how emotionally a text is written. Lastly, the word count of each title, description and story is calculated. Figure 4 plots these features against the funding amount raised. The bell-shaped distribution suggests that these features might be indicative of funding success, and that a sweet spot of moderate sentiment, emotionality, and text length lead to best results compared to extremes.

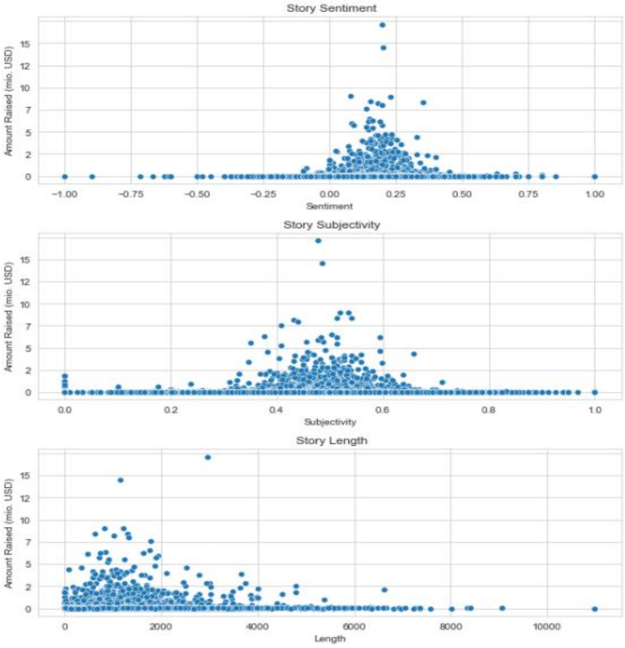

*Figure 4: Text Features from Campaign Story*

## B. *Construction of model*

To investigate the predictive power of the campaign title, description, story and geography, as outlined in research question two and three, a random forest classifier is trained on the data to solve the binary classification problem of predicting funding success. The advantage of a random forest model is that it is can easily deal with categorical data and that it is able to learn complex non-linear decision boundaries [5]. Thus, the categorical variables do not need to be one hot encoded. Instead, each category was assigned a numeric value. Observations, where the funding goal was could not be inferred, because they raised 0% of their goal were imputed with the median funding goal. This was done to prevent systematically discarding of unsuccessful campaigns.

Figure 5 shows the correlation matrix of all features used in the model, which gives a first insight into their relation with funding success.

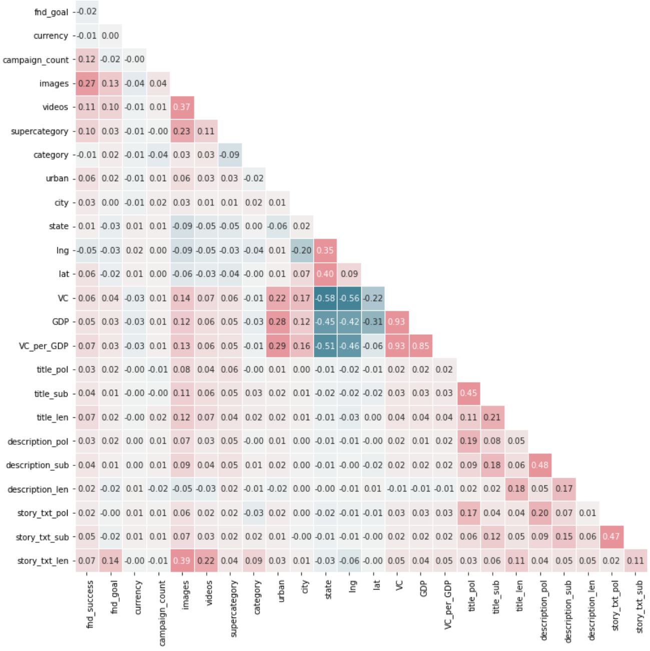

*Figure 5: Correlation Matrix Random Forest Features*

The model is trained on 80% of the total data. The other 20% of observations constitute the test set. 30 iterations of Bayesian optimization were used to tune the maximum number of estimators, the tree depth as well as the number of features and samples used in each tree. To prevent overfitting 3-fold cross-validation was used at each iteration. The best hyperparameters consisted of deep trees with depth 1000, and a high number of 2000 learners.

## C. *Validation of results*

To validate the results on the model, the model's performance is tested on the unseen data from test set. The model achieved an accuracy of 83%. Given that the test data contained 79% unsuccessful campaigns, this constitutes a 4% increase compared to the 79% baseline accuracy that could be achieved by only predicting "no success"Additionally, the model achieved high precision of 70%, but seemingly at the cost of relatively low recall of only 37%. The low sensitivity of the model is illustrated by the confusion matrix below.

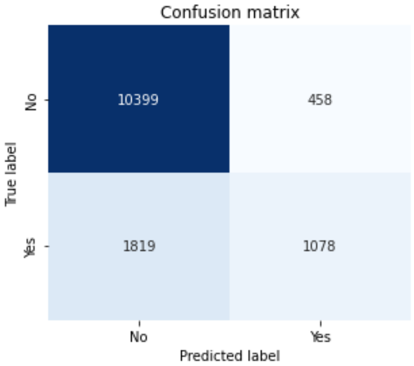

*Figure 6: Confusion Matrix Random Forest*

## V. Findings, reflections and further work

### A. Findings and reflections

Figure 7 shows the total funds raised for each category and colors them based on the corresponding sector.

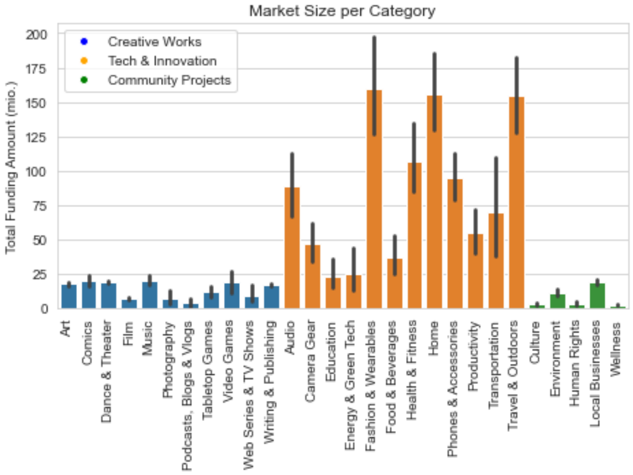

*Figure 7: Market Size per Category*

It becomes evident that the "Tech & Innovation" sector dominates the others in terms of size. Not only does it contain the most categories, but even its smallest category "Education" raised more funds than any other non-tech related category. Especially the "Fashion & Wearables", "Home" and "Travel & Outdoors" category attract a lot of capital, with market sizes north of $150 million each.

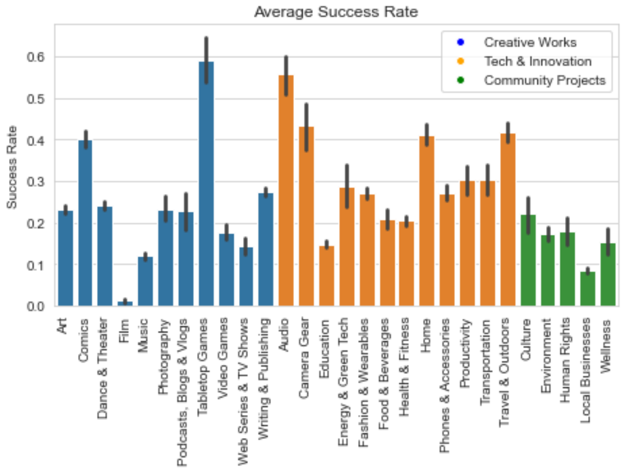

*Figure 8: Average Success Rate per Category*

Unlike for market size, there is not one sector that clearly has the most successful campaigns. The success rates are more dependent on the individual category. For example, both the most and least successful category, i.e., "Tabletop Games" and "Film", belong to the same sector "Creative Works".

Thus, to summarize with regard to research question one, both market size and seem to depend on the campaign's category. However, funding success seems to be influenced stronger by the individual category, whereas market size depends more strongly on the campaign sector.

Figure 9 shows the feature importance of the random forest model, which is used to gain insight into the predictive power of the geographic and campaign description related features.

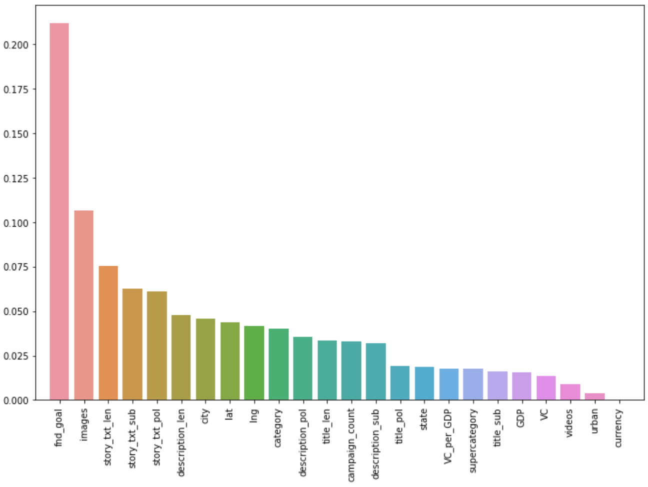

*Figure 9: Feature Importance Random Forest*

The funding goal is the most important variable to predict success, even though the linear correlation was not very strong (see figure 5). This suggests that the funding goal becomes an important feature not by itself, but when considering it in combination with other features like e.g. category, location and story text.

The next five most important features are all related to the presentation of the campaign through images and adequate subjectivity and sentiment. As suggested by figure 4 in the analysis part, there seems to be a certain range of values for story length, sentiment, and subjectivity that yield optimal results. Interestingly, the story features – not title or description – are the most important to the model. Potentially this is because the story is the most robust base to calculate these features on, as the longer wordcount reduces noise in the sentiment and subjectivity calculations.

These results show that there are indeed patterns in the campaign texts that predict success.

For geographic features, the city has the biggest impact on funding success, directly followed by longitude and latitude values. The high tree depth and the high number of learners seem to be able to uncover non-linear patterns in the geographic variable that predict funding success. Interestingly, the additional economic data on VC investments and GPD only play a minor role for the model, in spite of the relatively high correlation with funding success seen in figure 5. Potentially, this is because the city and coordinates provide a much higher resolution on the geographic dependencies that influence funding success and make aggregated data on a state level redundant. Thus, the research question two can also be answered positively, since geography seems to play a significant role in funding success.

### B. Further work

This analysis main limitation of this analysis is that causal effects were not investigated. This work has identified some important features that are relevant for the prediction of funding success. However, the causal relationship was not investigated. Thus, future studies could be to explore the causes-effect relationships that cause certain, categories, geographies and texts to be optimal for a successful campaign.

## REFERENCES

[1] Etter, V., Grossglauser, M., & Thiran, P. (2013, October). Launch hard or go home! Predicting the success of Kickstarter campaigns. In Proceedings of the first ACM conference on Online social networks (pp. 177-182).

[2] Gallemore, C., Nielsen, K. R., & Jespersen, K. (2019). The uneven geography of crowdfunding success: Spatial capital on Indiegogo. Environment and Planning A: Economy and Space, 51(6), 1389-1406.

[3] Lee, S., Lee, K., & Kim, H. C. (2018, October). Content-based success prediction of crowdfunding campaigns: A deep learning approach. In Companion of the 2018 ACM conference on computer supported cooperative work and social computing (pp. 193-196).

[4] National Science Board. "Venture Capital Disbursed per $1 Million of Gross Domestic Product." Science and Engineering Indicators: State Indicators. Alexandria, VA: National Science Foundation. https://ncses.nsf.gov/indicators/states/indicator/venture-capital-per-1-million-state-gdp. Accessed on 4.12.2021.

[5] Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer

| **WORD COUNTS*** | |
|---|---|
| Abstract | 154/150 |
| Introduction | 193/300 |
| Analytical questions and data | 247/300 |
| Data (Materials) | 289/300 |
| Analysis | 974/1000 |
| Findings, reflections and further work | 548/600 |

* not including headings and figure captions