Success Indicators of Crowdfunding Projects

Lukas Vlcek

Summary

Online crowdfunding platforms, such as Kickstarter and Indiegogo, have become popular places for individuals and companies to raise money for their projects and ventures from small investors. When deciding whether it is worth starting a funding campaign or investing money in manufacturing a novel product, it is important to evaluate carefully the expected success or failure of the project. To facilitate such considerations, it may be helpful to mine historical data on past projects to gain insights into the factors that determine success. In the proposed work I will analyze the available records from the two aforementioned platforms to find positive and negative indicators, and use machine learning algorithms to build a model for predicting the success of newly submitted projects. This information should be of interest to both startups and investors.

Background

The two major crowdfunding platforms, *Kickstarter* and *Indiegogo*, have together mediated the investment of billions of dollars to more than 200,000 successful projects. While there is a large overlap between the types of projects offered on the two platforms, there are some differences that should be considered in the proposed analysis. The websites provide various statistics based on their records, and mutual comparisons can be found on dedicated websites, such as at thecrowdfundingformula.com:

- Kickstarter offers 15 categories, which are more focused on arts and gadgets. The projects are only funded if the full amount is raised (all or nothing), and the overall success rate is around 36%.
- Indiegogo offers 24 categories, is somewhat more popular with technology and design, and often includes small business, manufacturing, and more controversial ideas. It allows the option of fixed (all or nothing) or flexible funding, and the overall success rate stands at ~17%.

While similar general statistics, such as average success rates, are readily available, they may not be capturing all useful information that can be collected from the records. At the other extreme, judging the intrinsic value of a unique novel product often requires a substantial domain knowledge, which may be impossible to glean from the past records. However, there exists a wealth of data characterizing each of the hundreds of thousands of successful and failed projects, which may contain non-obvious indicators of success that may go beyond simple statistics and provide a more objective context for making informed decisions.

Goals and target audience

It is my goal in the proposed study to analyze the available historical records from the two major platforms, and develop a model for predicting the probability of project success based on their quantifiable characteristics. The analysis and predictive models will help both startups and investors make informed decisions whether to invest their time, effort, or money into a given project. Also, individuals looking for new ideas could use the results to identify promising areas and avoid those with lower chance of success.

Datasets

The primary source of data for Kickstarter and Indiegogo platforms will be obtained from online, monthly updated repositories collected by a web scraping company Web Robots (webrobots.io), with data available in JSON and CSV formats.

1. Kickstarter data (4/2014 - present): Each item is characterized by 37 properties, including dates, deadlines, funding goals, and invested funds.
2. Indiegogo data (5/2016 - present): Each item is characterized by 22 properties similar to the Kickstarter dataset.

Each dataset contains more than 100,000 records of independent projects, which should be sufficient for statistical learning, but a closer inspection will be needed to see whether this is true for all project categories.

Web scraping could be used to collect additional data with significant information value, such as marketing efforts in the form of favorable articles, independent reviews, or responses under related video presentation (e.g., YouTube). However, this may be beyond the scope of the currently proposed work.

Approach

In the initial step, I will process the raw data and perform exploratory analysis to gain basic insights into the size of different categories and correlations, and to understand differences between the definition of similar categories on Kickstarter and Indiegogo. The results will help me in choosing suitable project characteristics (predictor variables), which may include some of the pre-existing characteristics, but also new ones, such as keywords in the project description or its length. The number and specific choice of the predictor variables as well as refinement of the target variables will depend on the quality and quantity of data for each of these variables.

Since the available data contain information about the target variables, *i.e.*, amounts raised and the success of funding campaigns, I will use supervised learning algorithms to train the predictive models. Simple binary classification based on decision trees can be used to predict two categorical (binary) variables: (i) a given proposal will achieve its funding goals, and (ii) a funded proposal will succeed in achieving its goals. However, because predicting the probability of success provides more detailed information I will also use logistic regression. Additional predictions of project delayed delivery can be included. For each problem, I will optimize a pair of models trained on the data from the two crowdfunding platforms, so that comparisons can be made.

As an exercise with possible new insights, I will employ unsupervised learning algorithm (PCA) to perform the analysis of correlations between different characteristics and reduce dimensionality of the problem by designing lower number of new, more informative descriptors (predictor variables).

The final discussion of the model predictions will include interpretation of the results, assessment of model strengths and weaknesses, and a consideration of identified correlations vs. causal relations between project characteristics and its success.

The data analysis and models will be implemented using Python with libraries for data wrangling (pandas), machine learning (scikit-learn), and plotting (Matplotlib/Seaborn).

Deliverables

The results will be presented in the form of a written report and presentation slides. The project files, including sources for models and references to the online datasets, will be made available as open source on Github. In case interesting insights and predictive models are gained from this study, it may be worth presenting it in the form of a blog post for a wider audience..