RESEARCH PAPER:

KICKSTARTER PROJECTS

DATA ANALYSIS

Seminar Paper of Business Informatics 2

Submitted: 30th September 2023

By: Nigar Salmanzade

Reviewer: Prof. Dr. Frank Köster
Christian Janßen, M. Sc., Viktor Dmitriyev

Carl von Ossietzky University Oldenburg

Table of Contents

Appendix - Country abbreviations

AT-Austria

AU-Australia

BE-Belgium

CA-Canada

CH-Switzerland

DE-Germany

DK-Denmark

ES-Spain

FR- France

GB- United Kingdom

HK-Hong Kong

IE-Ireland

IT- Italy

JP-Japan

LU-Luxembourg

MX-Mexico

NL-Netherlands

NO-Norway

NZ- New Zealand

SE- Sweden

SG- Singapore

US-The United States

1. Introduction

A new method for individuals and teams to ask for financial assistance from a global audience is through crowdfunding, Crowdfunding is the online appeal for resources from a distributed audience frequently in exchange for a reward. Crowdfunding, in contrast to conventional fundraising techniques like requesting money from banks or institutions, enables creators—people who want resources—to directly request money from donors—people who provide resources—through internet platforms. One of the most popular websites for creative individuals to crowdfund projects is Kickstarter. Here, creators may discuss their fresh ideas for creative projects with the groups who will support them financially. Kickstarter has supported a wide range of projects and is a prominent crowdfunding site in the online services area. Since the platform's debut (2009), 21 million individuals have supported projects, raising a total of $3,430,261,148 and 326,133 of those projects have been financed.

Projects on Kickstarter come in all shapes and sizes, and they cover a wide range of topics, including the arts, comics, dance, design, fashion, cinema, cuisine, gaming, music, photography, publishing, technology, theatre, science, and services.

The amounts raised range from a few dollars to more than ten million dollars. Kickstarter projects have a success rate of 35 percent.

In our paper we analyse the Kickstarter project data from 2018. We follow the methodology of the Cross-industry standard process for data mining (CRISP-DM) and go through the steps to determine if a project is successful. This analysis will provide guidance to the project owners to make their projects more attractive for the backers. In the sections below we first provide a business understanding of the data, then move on to Data Understanding and Data Pre-processing, Perform Exploratory Data Analysis and finally build a Model and fine-tune the model from the results.

2. Business Understanding

There are two types of users who will be interested in the crowdfunding data. One, the Project owners and the other being companies. For crowdfunding companies, the data can be used to filter the projects. Noe from the project owner’s perspective, We are

currently living in a consumerist society. Project owners and the number of projects is growing every second. Faced with this reality, individuals need to find avenues to make their project funding successful. One would think project or product quality is enough for securing the funding. In our analysis, we show that various variables like category of the project, features like dates, name of the project affect the chances of it being a successful project. Our main objective in this paper is to predict if a project will be successful in achieving its funding goal with the help of explanatory variables. We do this by using Machine Learning Algorithms like logistic regression, Random Forest. While the above-mentioned being the main hypothesis we also provide insights on some interesting questions which might be brewing in the project owner's minds.

1. Which category of projects are more likely to be successful?
2. Does Date, Month, Quarter, Weekday make an impact on project funding?
3. What is the Average funding required for different categories of successful projects?
4. What are the keywords in a project name and the appropriate length of the name?
5. What is the average funding for the projects?
6. Success rate of different categories in different countries?

3. Data Understanding

There is one data set of Kickstarter projects from the date of April 21st, 2009, to January - 2nd, 2018. In this part we will inspect the data and perform exploratory data analysis. The data has 15 variables and 378,661 rows of values. Most of the variables in the data are symbolic and objective in nature, there are Date Time variables. Some of these variables are used to create new variables like the length of the words in the project name, the number of weekdays between the projects. Out of the few numerical variables we must drop the number of backers and the United States Dollar (USD) pledged because these are directly correlated to the project being successful.

The Kickstarter platform data can be divided into the following categories.

1. Project name - The name variable in the data is a string containing the names of the projects. In the paper, we turn these strings into a vector of values and convert them into a sparse matrix.

2. Categorical variables - Kickstarter projects are divided into categories and subcategories to ease the process of searching for projects. These attributes will be used for checking successfully funded categories. The variables currency and country show the origin country of the project and the currency of the project funding.

3. Datetime variables - launched and deadline columns in the data show the date of project launch and deadline date of the project. These variables will be changed to a datetime format for further analysis. Variables like a quarter, month and year are parsed from these variables.

4. Numerical variables - United States Dollar (USD) pledged i.e., funds raised, funding goal and backers are the numerical variables in the dataset.

3.1. Verifying Data Quality

3.1.1. Missing Values

The variable name contains four missing values and the variable USD pledged contains 3797 missing values. However, this variable is the pledged amount in USD converted by Kickstarter, and we use the usd_pledged_real which has all the values. Hence, Missing values are not a problem in this case.

3.1.2. Outliers

In the data, we check the outliers for the variables usd_pledged_real and usd_goal_real since they represent the currency values and are among the few numerical variables in the dataset. The number of outliers for usd_pledged_real and usd_goal_real are 50,578 and 45,508 respectively.

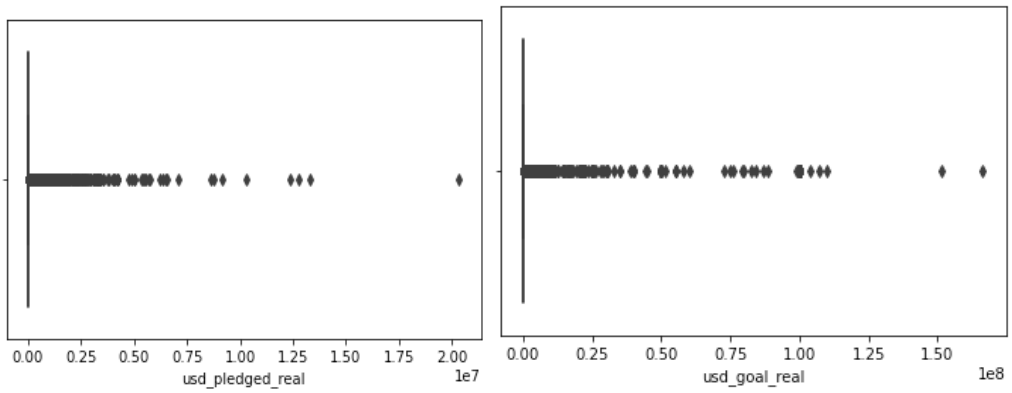


Figure 1. Plots of Outlier

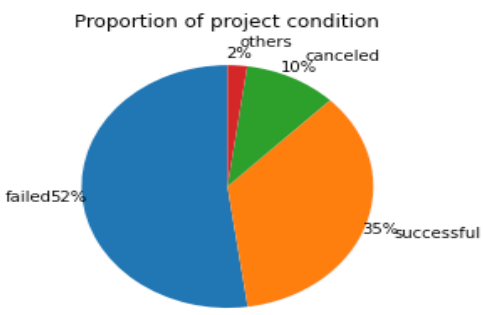
3.1.3. Data Integrity

In this dataset, we have country values called N,0" which is inconsistent, but it is not a problem as the state of the project for these values is undefined. In our study we are only concerned about either successful or failed projects. Overall, the quality of the dataset is very good. So, it is suitable for accurate data analysis.

4. Exploratory Data Analysis

4.1. State of the Projects

Among all the Projects there is a success rate of 35 percent which shows great potential for crowdfunding projects. The number of successful projects is 133,956 and failed is 197,719. In the table, Undefined or Cancelled can mean that it is possible that the campaign has not been launched yet, No longer active because they violated the terms of service. It can also mean incomplete data.



Project state	Number	Share
failed	197,719	52.22%
successful	133,956	35.38%
canceled	38,779	10.24%
undefined	3,562	0.94%
live	2,799	0.74%
suspended	1,846	0.49%

Figure 2. Pie Chart and Table

4.2. Money Estimates

- Average values for Funding goal, Amount pledged and Number of backers
- The average Kickstarter campaign has a goal of 49080 USD, has 106 backers and an average campaign has a pledge of 9683 USD. Average estimates for successful and failed projects are also given below. The goal and pledged amount are in United States dollars.
- Average contribution by a backer.
- Average contribution by a backer is 76 USD, this measure has been attained by first dropping rows where there are zero backers and then making a new column of average contribution.
- Highest backed and Least backed Product.
- Pebble Time Smartwatch was the highest backed product with the amount raised being 20,338,986 USD. Their goal was 500,000 USD.
- Allenby leather bags was the least backed product. An individual who backed them has invested 1.01 USD and their goal was 10,066 USD. This product is the least backed among campaigns where the goal was greater than 1 USD.

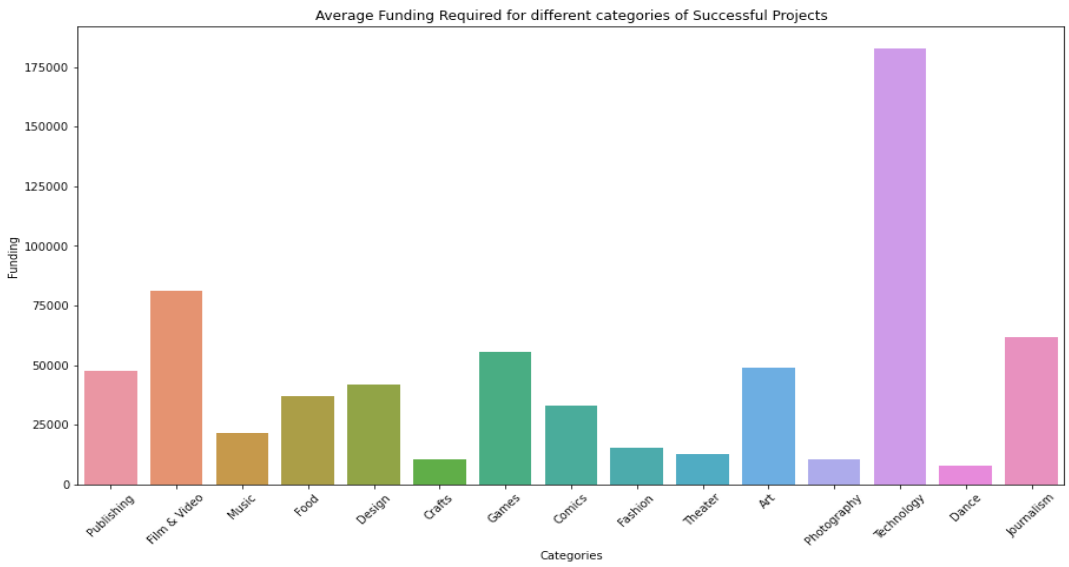


Figure 3. Category Bar Chart

4.3. Country and Project category evaluation

- Since Kickstarter is an American corporation and America being the hub of capitalism it is no shock that around 78 percent of the projects are from America. It is followed by Great Britain and Canada.
- In the category section. Film & Video has the highest share of the projects at 16.79 percent and 63,585 projects. Followed by Music and Publishing. It is surprising to see that Technology, Food and Fashion account only for 21 percent of the share.
- In the table below we calculate the success rate of projects by their categories and countries. The index denotes the country names, abbreviations of the countries are given in the appendix. The null values in the table represent that there are no particular projects of the category originating from the particular country. The United States hosts the highest number of successful projects. Out of all the categories Comics, Theatre and Music stand out. Even though Dance has a success ratio of 62 percent because it has only 3767 entries, it cannot be considered significant.

4.4. Datetime Variable Analysis

4.2. Money Estimate

The data has two Datetime variables namely launched and deadline. These dates were in the format of unix time, the values were parsed to get year, month, quarter, weekdays between launch date and deadline date. The analysis is presented below.

1. Years

Kickstarter started in 2009 with 1179 projects. In the first year itself the number of projects grew by over 8 times to 9577. The number of projects reached a peak of 65,272 in the year 2015.

2. Month and Quarter

In the graph below we plot the number of successful projects filtered by quarter and month. We can clearly see that the project is less likely to succeed in the first quarter of

the year. Individuals should try and avoid releasing the projects for funding in these months.

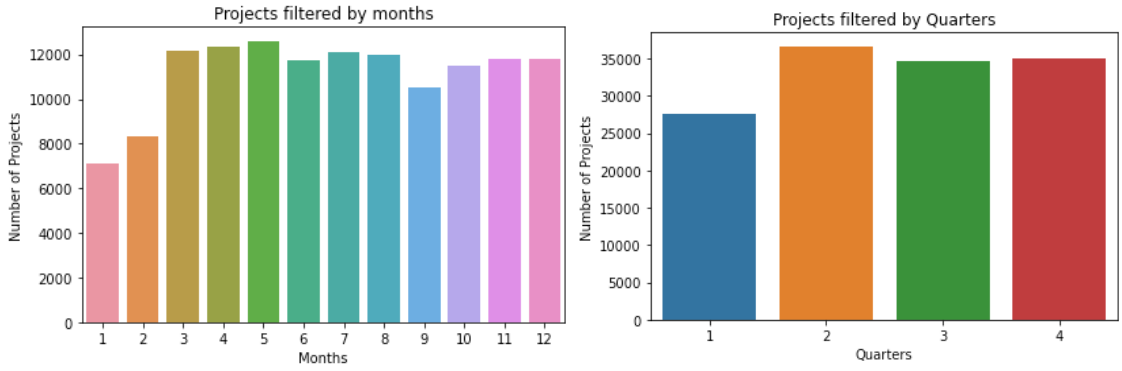


Figure 4. Category Bar Charts

3. Weekday launched and Number of Weekdays

The graph shows that most of the projects are launched on the weekdays, and we plot a graph showing that the number of weekdays between projects is not as important as one would have hoped.

4. Project name Analysis

A new variable word length was created. Word length denotes the length of the project name. The mean length of the words is 34. The max length is 85. Most of the projects have a word length between 21 and 48.

5. Data Pre-processing

The Main Hypothesis of this project is the prediction of project success. We first filter the state variable and use only successful or failed projects for our analysis. We then move on to dropping unnecessary variables. The variables are the goal, pledgUSD usd pledged, and usd_pledged_real, backers are dropped because of their close correlation to the dependent variable state. The deadline dates parsed are dropped because of their close correlation to launched dates. The category is dropped in favour of main_category. Now, the model is left with categorical and numerical variables.

5.1. Dealing with Numerical Values

The numerical variables are usd_goal_real and weekday count. Since this involves currency and numbers. The numerical values are scaled using sci-kit learn library scaling. Here, we preferred scaling to normalisation because normalisation makes the distribution normal which is not needed in our case and scaling is used to compare numerical variables on an equal footing.

5.2. Dealing with Categorical Values

For the categorical values in the data, we used the dummy variable method to create dummy variables for all the categorical values in our data set. This transformation is important because most of machine learning models only read numerical data.

The data after being scaled and creating dummy variables have 68 variables and 331675 values. The data is then split into train and test data with the test data size being 30 percent of the original data set.

5.2.1. Variables Used for Model Prediction

Goal of the project in US Dollars, Number of Weekdays between launched and deadline date, main category, year launched, month launched, country and launched weekday.

6. Modelling

In our model we have a categorical variable state of the project as the dependent variable which we try to predict. The idea behind our prediction is given the goal and country and weekdays between launch date and deadline date. We can make an accurate guess if a project is a success or a failure.

The predictor variable is a categorical variable. So, Linear Regression is ruled out. We can use both Supervised and Unsupervised learning methods on the data. In our analysis we stick to supervised learning methods. An ensemble of models namely Logistic Regression, K Nearest Neighbours, Random Forest, Bagging and Boosting classifiers, Natural Language Processing are used for modelling the data. The use of these

multitude of methods is possible because the size of the data being small in nature. In the section below introduction to the machine learning algorithms is provided.
Logistic Regression estimates the probability of an event occurring when supplied with independent variables. It is the most common model and has high interoperability. K Nearest Neighbours classifier creates a similarity index and then classifies the data point based on similarity. Random Forest, Boosting and Bagging are ensemble methods. They combine several predictive models to arrive at a better prediction rate or accuracy. The ensemble methods we used in the data have a base estimator as a decision tree. In which the best variable is taken at each level, and it is used to make a cut at the point where the error is the lowest. In our case it is the Gini error because we are using a decision tree classifier.

We also make a model for predicting the success using Natural Language Processing (NLP). NLP translates human readable language into machine readable language. We first remove punctuation marks and stop words from the project name and then make a count vector out of the project names. This process involves conversion of text into a sparse matrix of token counts. After this, we use Term frequency and Inverse Document frequency (Tf-idf) for transforming our count vector.
Term frequency measures how frequently a term appears in the corpus. In our data set the corpus being the project name. Inverse Document frequency measures how important a term is by providing weights. We use the transformed sparse matrix for the prediction using logistic regression.

7. Evaluation

In our model evaluation we used the metric accuracy score as the metric to determine the error. Accuracy score is the fraction of instances where the classification was predicted correctly. It is a good estimate for an error with a classification model as a dependent variable. We divide the data into a training set and test set. The model is first fit using the training data and then training accuracy score and test accuracy score were evaluated.

Out of all the Algorithms we used to fir the model Logistic regression and Ada Boost performed very well. They have a train and test score of 0.65 and 0.66. Random Forest Classifier has a train score 0.97 which is almost perfect with low bias but it's test score is 0.63 which id low. It is often the case that a model with low interpretability outperforms a complex model.

In the NLP model we get a train score of 0.734 and a test score of 0.65 this shows that almost all the models have the same accuracy score, but Adaboost does a bit better. In our analysis we found that having the category name in the name of the project is quite common and most of the words contain the category names.

In the next step we further fine-tune the model using cross-validation. We use cross-validation to find the best parameters for logistic regression and Random Forest and estimate the error. The best parameter for logistic regression is Inverse of regularization strength (C) equal tpo100 a higher value denotes the parameter not being regularized. The best parameters for Random Forest are minimum sample leaf as 5, minimum sample split as 6, number of estimators as 200 and maximum depth of 7. The accuracy for the Random Forest increases from 0.63 to 0,64 whereas the accuracy of the Logistic regression does not increase.

In our NLP model we checked for the most common words in the project name column. Most of the common words are the subcategory values. The top 5 words being Project, Album, New, Film, Book.

Feature Importance.

The feature which is most important is the usd_goal_real with a weightage of 0.36, it is followed by weekday_count which accounts for 0.16 of the change in dependent variable.

8. Conclusion

In our analysis we have seen that having certain features can increase the chances of the project being successful. It is important to set an optimal goal. Setting a high goal is not going to help a project in securing funding. Having higher number of weekdays between the launched date and the deadline date increases the chances of reaching

funding goal. In the case of category and subcategory the chances of getting funded are high in case of a popular category. Through our applied machine learning algorithms, we got an accuracy score of 0.66, this could be improved trying different combinations. We could remove certain countries from the data to make the variable more significant. Also, additional data could be integrated like innovation index to increase the accuracy score.