Universiteit
Antwerpen

Academic year 2021/22

Faculty of Business and Economics

Data Engineering Project

Data Engineering

Prof. xxx

Group members:

xxx

xxx

1. Introduction and problem understanding

We as students from the Faculty of Business and economics at the University of Antwerp were asked to carry out a group project for the course Data Engineering. The dataset that we were given was one of Kaggle.com that contained data on all kinds of aspects of crowdfunding campaigns that were launched on the popular website Kickstarter.com. The crowdfunding platform allows entrepreneurs and creators to pitch their ideas to a large group of potential investors with the goal of gathering funds in order to 'kickstart' their projects. A crowdfunding campaign via Kickstarter.com works as follows: Every project creator sets their project's funding goal, deadline and optionally reward for investors. If people like the project, they can pledge money to make it happen. If the project succeeds in reaching its funding goal, all backers' credit cards are charged when time expires. Funding on Kickstarter is all-or-nothing which means that if the project falls short of its funding goal, no one is charged.[1] Using this information, we could therefore classify a campaign as successful if the amount pledged exceeds the predefined monetary goal.

The goal of this project was to gain insight into what factors constitute to making a campaign successful or not. Such an analysis could be useful for investors of future crowdfunding campaigns. For the machine learning part, we therefore chose to carry out a binary classification to see if we could predict whether a campaign would become successful or not. In what follows we'll first provide an exploratory analysis followed by the data preparation and machine learning part.

2. Exploratory analysis

Summary statistics

To start of the exploratory data analysis, it is interesting to give an extensive overview of numerous summary statistics that can give an idea of what kind of data is being dealt with as well as some insights into the world of crowdfunding projects via Kickstarter.com. The table below gives 8 of the most important statistics which often also gives the distinction between successful and non-successful projects.

Statistic		Value
Number of campaigns	Total	372,780[2] campaigns
	Successful	133,851
	Failed	238,929
Percentage of successful projects		35,91%
Total amount of money pledged	Total	$3,402,632,648.09
	Successful	$3,033,664,091.26
	Failed	$368,968,556.82
	Total	$9,127.89

[1] Explanation from www.kickstarter.com

[2] After data cleaning, not including projects that were still 'live'

Average amount pledged per campaign	Successful	$22,664.48
	Failed	$1,544.30
Average amount of backers per campaign	Total	106,77 backers
	Successful	264.13 backers
	Failed	18.62 backers
Average amount pledged per backer	Total	$75.66 per backer
	Successful	$91.12 per backer
	Failed	$64.60 per backer
Average duration of campaign	Total	33.61 days
	Successful	31.59 days
	Failed	34.74 days
Average goal of campaign	Total	$45,658.35
	Successful	$ 9,535.70
	Failed	$65,894.72

Table 1: Summary statistics of dataset

Although all the summary statistics in the column above provide us with some useful information, there are a few that are especially interesting. Take for example the average duration of a campaign: one would expect that a campaign that lasts longer in general would have a higher chance of succeeding. The data however tells us otherwise as successful campaigns on average have a shorter duration than failed campaigns.

Furthermore, it is also intereting to note that successful campaigns on average set their goal much lower (nearly 5 times lower) than unsuccessful campaigns. Also, successful campaigns have a much higher amount pledged to them on average (nearly 15 times more) than unsuccessful campaigns. We could therefore assume that investors are much more likely to back a project and pledge a higher amount if the campaign sets itself realistic goals that are attainable. Unrealistic campaigns, with goals that are perceived as to high, will probably dissuade investors from investing.

Succesrate categorized by country

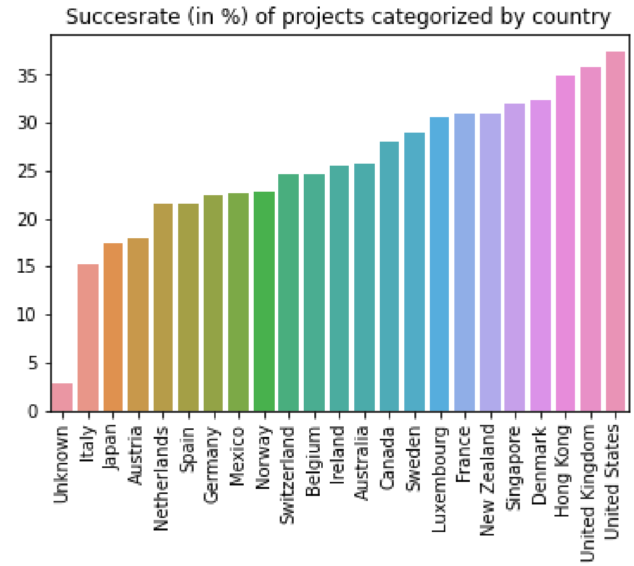


Figure 1: success rate (in %) of projects categorized by country

One interesting insight that we gained from performing our exploratory analysis was by analyzing the success rate of campaigns by country. Here we found that the country that had the highest rate of success was the United States. With a rate of a little bit over 35% it was more than double as likely to be successful than Italy, the country with the lowest rate. This could partially be explained by the fact that Italy's biggest main category was technology (20%) whilst for the US only 7% were categorized as technology projects.

Success rate categorized by main category[3]

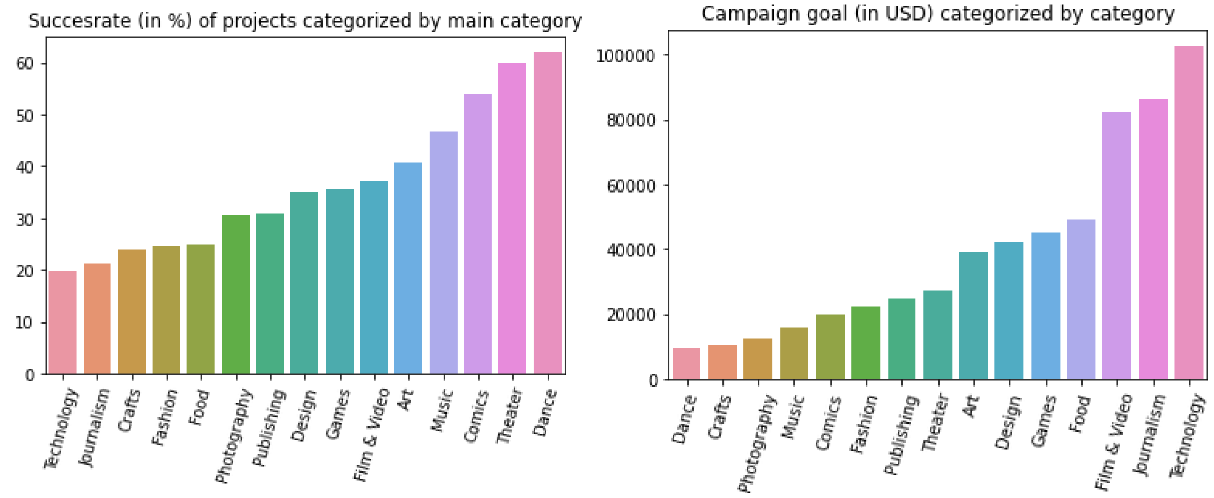


Figure 2, 3: success rate (in %) of projects categorized by main category, campaign goal (in USD) categorized by category

Each campaign had a main category that got assigned to it. Consequently, it was also interesting to analyze what type of projects were most likely to succeed. Looking at figure 2 above, we can conclude that projects in the category 'Dance' were the most likely to be successful whilst projects in the category of 'Technology' were least likely to be successful. This might seem counterintuitive at first, as the stock market in the last couple of years has highly valuated technology initiatives. However. this could be explained by the fact that the campaign goals for technology initiatives are on average much higher than for example dance initiatives, and as we saw in the previous section, campaigns that are perceived as attainable are much

[3] In the apendix a pie-chart can be found with the distribution of the various main categories

more likely to be successful. Also, there were much fewer dance initiatives than technology initiatives (1.3% vs 8.6%) which could also explain the situation.

Relation between success rate and date

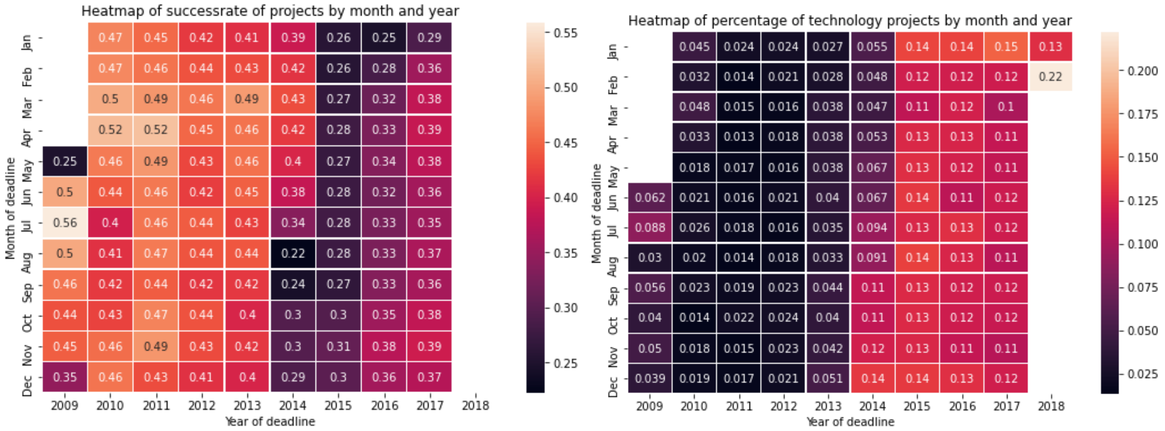


Figure 3, 4: Succes rate categorized by month and year, percentage of technology projects categorized by month and year

The final graph that we examined was one that plotted the success rate of campaigns in a certain month and year (figure 3). Here, the lighter the tile gets, the higher the success rate is. The goal was to evaluate whether some months or years had a higher success rate than others. With regards to the different months, we can observe that there aren't huge fluctuations. Seasonal effects are therefore not really a factor that influences the success rate. If we look at the years however, we can clearly see that from august 2014 and onwards, there was a decline in the success rate which only recovered a little bit at the end of 2016. One explanation could be the rise in the percentage of technology campaigns, which as we saw before, had the lowest success rate. This rise from July 2014 onwards is clearly depicted in figure 4. Another explanation could also be that people lost interest in the platform at this given period which would have also led to a decline in successful campaigns.

3. Data Preparation

3.1 Checking data comprehensiveness and duplication

Before we preprocess any data, it is always a good practice to have some basic checks on the data. We checked the column names and whether there were any duplicated items in the data set. Also, we ensured whether there are any non-numerical values in the numerical attributes and whether there are any numerical values in the categorical attributes (i.e., "state", "currency" and "category").

There are no duplicated items in the data set provided or any misplacing of data types unless for the feature "launched" and "deadline".

3.2 Performing data parsing and creating timestamp features

The feature “launched” and “deadline” record when the crowdfunding campaigns started and when was the end. However, the data was not recorded in a proper time format. Therefore, we transformed the data into a proper time format (i.e., “datatime64[ns]”) for potential further use in machine learning by data parsing. Data parsing is the process of transforming data from one format to another format.

Then, we created three different features, “day”, “month” and “year” from the feature “launched”. It made us easier to group the data in terms of days, months or years for further exploratory analysis or data visualisation. Later on, we further used the time data to explore in exploratory analysis.

3.3 Classifying the feature “state” into 3 multivariate classes

In the feature “state”, there were initially 5 different classes. They were “failed”, “cancelled”, “successful”, “live” and “suspended”. We re-classified them into only three multivariate classes to better prepare the data for further analysis and machine learning. For the “successful” case, we labelled them “1” in a new feature “outcome”. For the “live” case with “usd_pledged_real” higher than “usd_goal_real”, we also labelled them “1” in the new feature “outcome”. But for the special “live” case with “usd_pledged_real” lower than “usd_goal_real”, because of the uncertain status, we labelled them “2” first in the new feature “outcome”. When working on the exploratory analysis or machine learning, we would consider how to treat this type of data again. For all those other cases like “failed”, “cancelled” and “suspended”, we all labelled them “0” in the new feature “outcome”.

3.4 Handling missing values

We counted the number of missing values and found out that the major missing values were from the features “state”, “country” and “usd pledged”. There were only 3,801 data instances that included missing values and accounted for 1% only of all observations (i.e., 378,661 observations). Therefore, we decided to drop the rows with missing values because it would not result in a serious data loss.

3.5 Performing Min-max scaling

After converting from the original currencies to US dollars, the features “usd_goal_real” and “usd_pledged_real” were produced for easier comparison. However, the range of the US dollars varied, and it would affect the weighting in the model when working on machine learning. We normalised and limited the range by using Min-max scaling. It is the simplest method of rescaling the range of features to scale the range in [0, 1] or [−1, 1]. We chose Min-max scaling instead of other methods like Mean normalisation or Z-score normalisation because Min-max scaling would not distort the distributions of the attributes.

4. Modeling and Evaluation

We start of our Machine Learning part with the csv file generated at the end of the pre-processing phase. Our question was the following: Can we predict whether a crowdfunding campaign on the Kickstarter platform will be successful or not.

4.1 Machine Learning specific pre-processing

Before training a model on a dataset, some work must be done. First, we dropped the features that are irrelevant and redundant like ‘ID’, ‘goal’ and ‘pledged’. Secondly, since we’re only interested in projects whose ‘state’ column is either ‘successful’ or ‘failed’, we dropped other values in the rows. Having further examined the data, we also found the ‘country feature of a small subset of rows has a value of ‘N,0’, which appears to be malformed, these rows were also excluded from subsequent modelling and testing. Next, we decided to create a new integer feature, ‘duration_in_days’, based on the difference in days between the two dates found in ‘deadline’ and ‘launched’, this tells us ‘Will a project raise more money if its fund-raising period lasts longer?’. After it we dropped both ‘deadline’ and ‘launched’ columns. Then we went on to encode categorical features, and transformed the features ‘main_category’, ‘country’, and ‘state from string representations into integer features. We encoded ‘successful’ to 1, and ‘failed’ to 0 for ‘state’. We applied one-hot encoding to ‘main_category’ and ‘country’ and created features such as ‘is_country_US’, ‘is_country_CA’, ‘is_category_Technology’, etc. There are 15 distinct values for ‘main_category’, and 22 for ‘country’. These 37 columns, combined with ‘state’, ‘backers’, ‘usd_goal_real’, and ‘duration_in_days’ lead to a total number of 41 features.

At last, we shuffled the data, and splited the data into training and test set. We decided to use 70% of the remaining data as training data, and 30% for testing. Our dataset is now ready for the data modeling phase. We trained different ML algorithms: K-Nearest Neighbor, Decision Trees, Random Forest, Support Vector Machine, Logistic Regression with stochastic gradient descent, Neural Network, Perceptron, Ada-boost, Bagging, Gradient boost.

4.2 Evaluation

We implemented the baseline model to predict the mode or majority label. The mode of the ‘state’ column for both training and test set is 0 (i.e., ‘failed’). For this baseline model, the accuracy is simply the percentage of failed projects among all projects in the test set, which is 59.9%. Since this is a supervised learning problem in which the correct labels are given, we chose to use accuracy and F1 score as the evaluation metrics.

We found the best accuracy and F1 scores are obtained on the testing data when Random Forest is employed. The result is an accuracy of approximately 0.94, and F1 score 0.95. From looking at the learned weights in decision trees and random forest, we found the feature that matters the most for predicting the funding outcome to be the number of backers, followed by the amount of funding the project tries to raise. The results make sense, since one may think the more backers a project has, the more popular it is, and thus it's more likely for the project to get the support it needs to be successful. This result also corroborates with the best-performing baseline model, where the only feature is backers.

5. Conclusion

The thing to try further is to use different features to improve the performance of the classifier. With more knowledge and experience, we will be able to employ these models more effectively and have clearer ideas on what to look for regarding grid search and hyper-parameter tuning.

When running the machine learning algorithms, at first, we tried the Spyder in Anaconda we learned to use in Data mining course last semester, but because of the data size, that didn't work well. We then used the Google Colab we learned to use in this course, all the things ran well, so it gave us a deeper understanding of the advantage of cloud computing.

6. References

Sebastian Raschka, Vahid Mirjalili(2019). Python Machine Learning, 3rd edition.

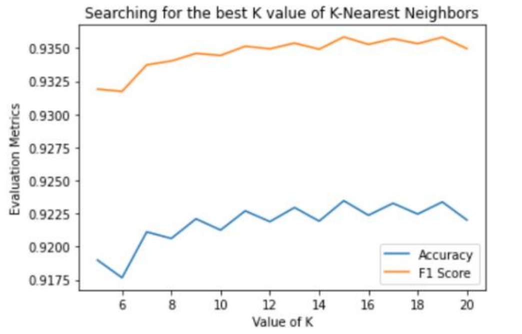
Léon Bottou(2012). Stochastic Gradient Descent Tricks.

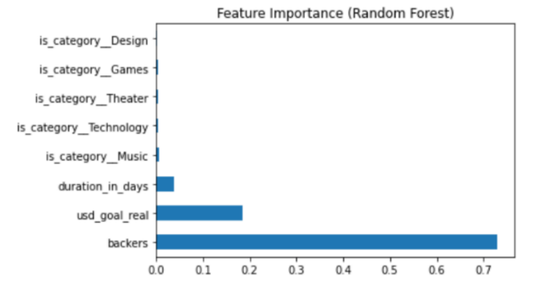
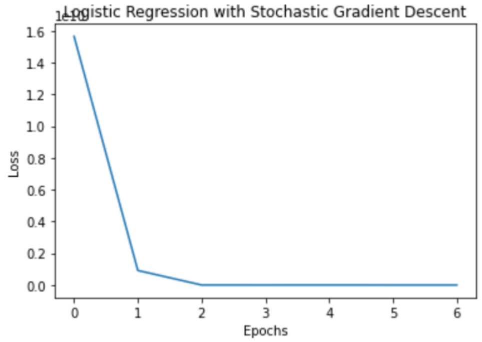
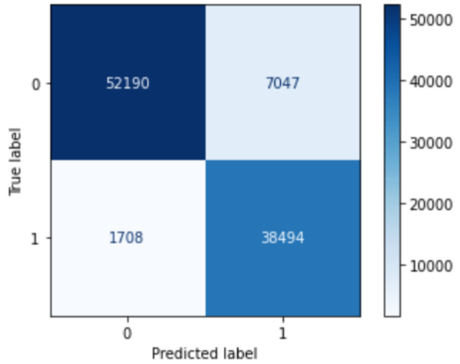
Aurélien Géron(2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition.

David Martens (2021). Data Mining Lecture Notes

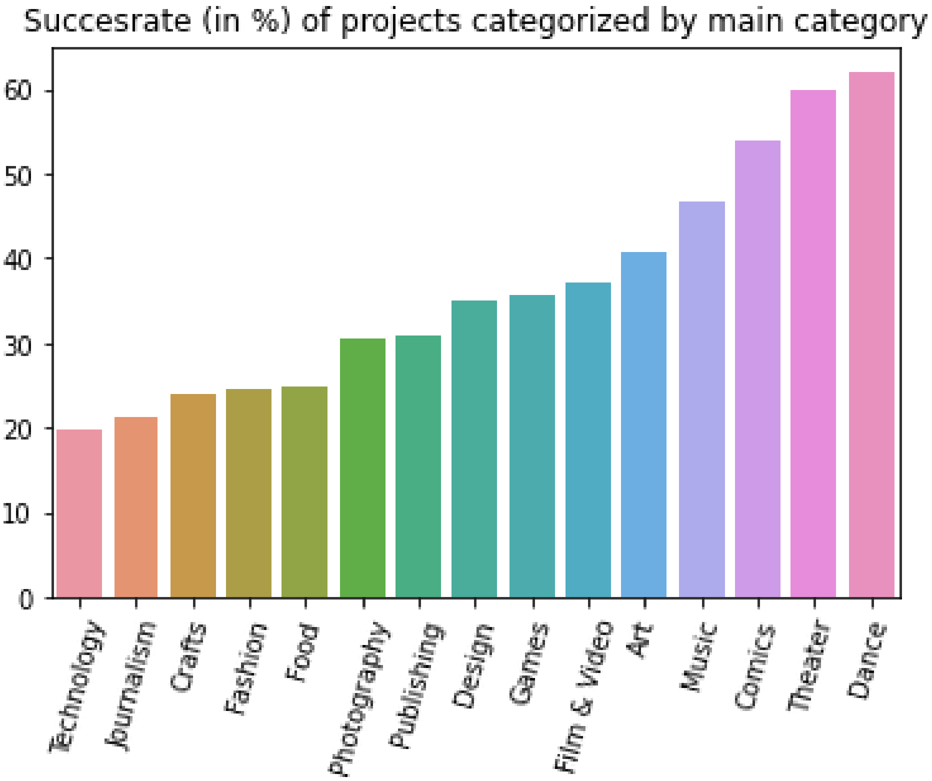
7. Appendix

7.1 Table with machine learning results

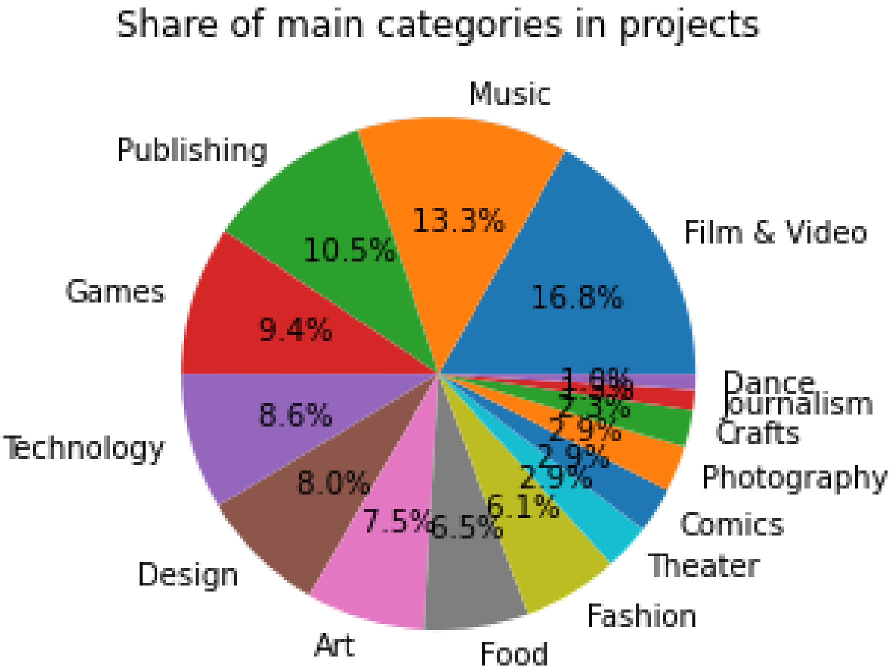
ML Algorithm	Accuracy	F1 score	Extra information
Baseline	0.599	/	
K-Nearest Neighbors with optimal k value: 15	0.923	0.936	
Decision Trees	0.929	0.940	

Random Forest (Best practice)	0.935	0.945	
SVM	0.627	0.460	
Logistic Regression with stochastic gradient descent	0.911	0.910	 
Neural Network	0.926	0.925	
Perceptron	0.869	0.870	
Ada-boost	0.933	0.930	
Bagging	0.932	0.940	
Gradient boost	0.934	0.944	

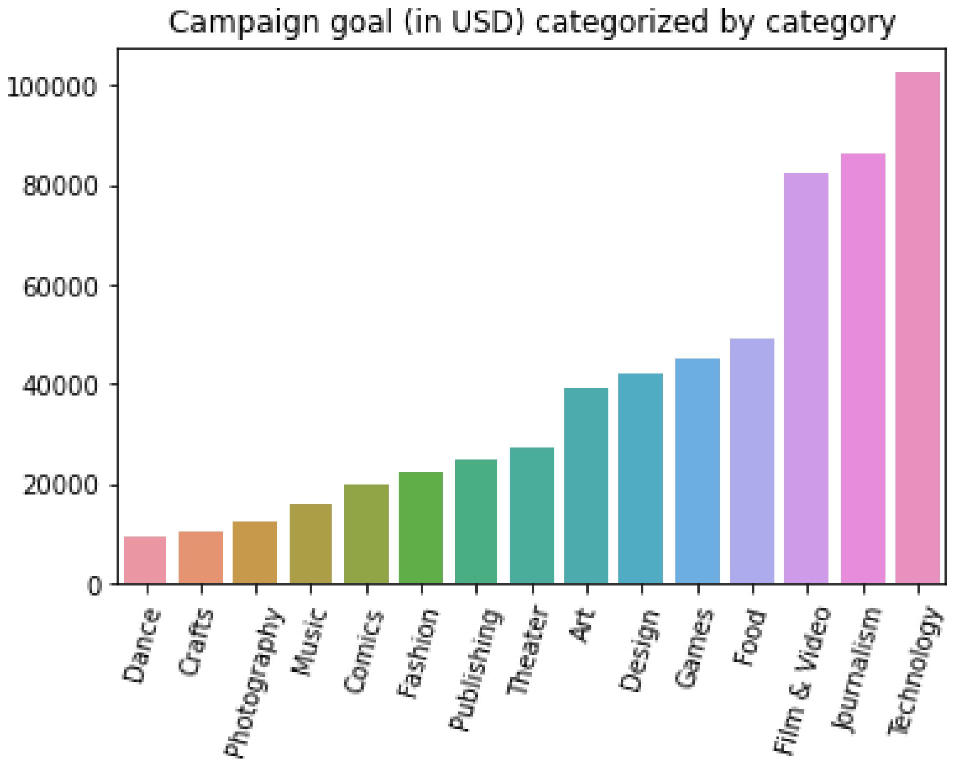
7.2 Figures

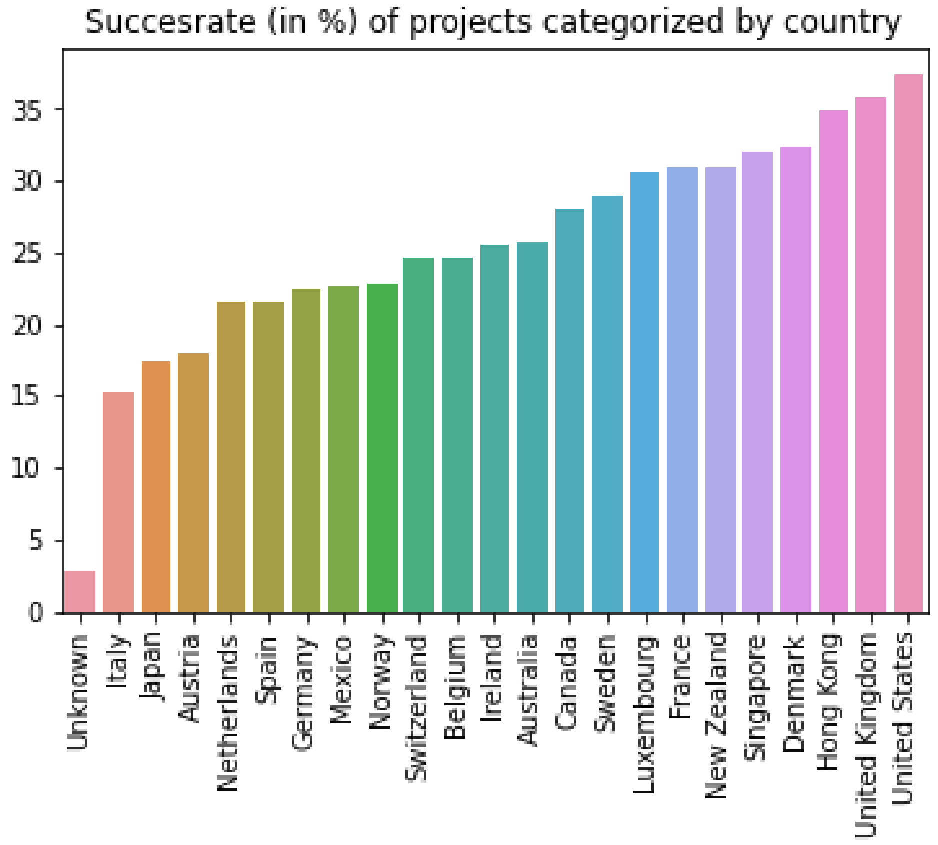


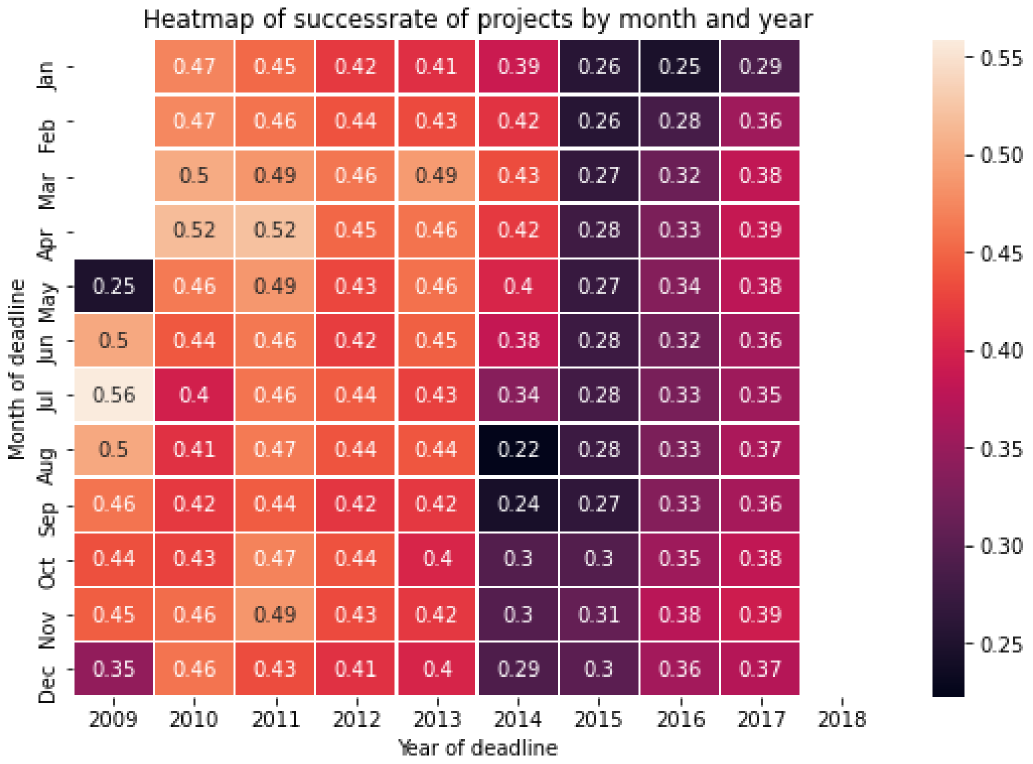
Succesrate (in %) of projects categorized by main category



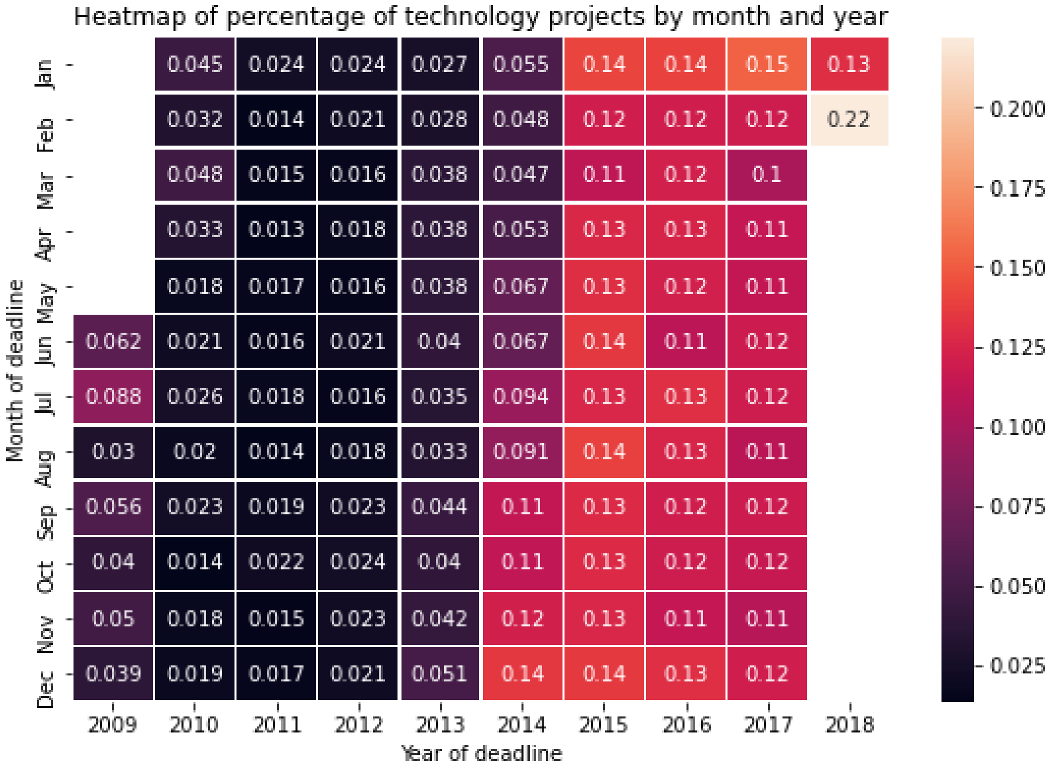
Share of main categories in projects


Campaign goal (in USD) categorized by category
100000
80000
60000
40000
20000
0
Dance
Crafts
Photography
Music
Comics
Fashion
Publishing
Theater
Art
Design
Games
Food
Film & Video
Journalism
Technology


Succesrate (in %) of projects categorized by country
35
30
25
20
15
10
5
0
Unknown
Italy
Japan
Austria
Netherlands
Spain
Germany
Mexico
Norway
Switzerland
Belgium
Ireland
Australia
Canada
Sweden
Luxembourg
France
New Zealand
Singapore
Denmark
Hong Kong
United Kingdom
United States



Heatmap of successrate of projects by month and year



Heatmap of percentage of technology projects by month and year