**DIGGING INTO KICKSTARTER DATA**

*Hasan Alperen Tosun Elif Kır Taha Bayaz Ufuk Seçilmiş*

In this paper we focus on how the creative ideas are funded on the crowdfunding platforms. For our purposes we will use an extensive dataset of Kickstarter projects and examine the structure of funding success for different categories of projects. Then we develop machine learning algorithms to predict the success of the projects. We used Decision Tree and Naive Bayes Classifiers and got 67% of training accuracy. Our work expands the literature in terms of predicting the success of the new entrants, which is a efficiency improving process.

**INTRODUCTION**

Is it possible to predict by using previous crowdfunding data and machine learning algorithms whether an idea is good enough to be funded? If we can, how powerful is the prediction? Even further, can we rely on a continuously evolving ML algorithm to decide on funding of the projects. These questions have power to influence the fundamentals of entrepreneurship if answered.

In this paper we are going to use the dataset of crowdfunding projects, which were advertised at the Kickstarter website, and we will at least try to find some answers for above-mentioned questions. Kickstarter.com is a website where people with ideas but without money go online and ask for donations. It is a public-benefit corporation that maintains a global crowdfunding platform focused on creativity and merchandising. This dataset consists of more than 300,000 projects which were launched at the website before the 01.02.2018. Within the dataset we have the information for the name of the project, the category it belongs, the main category it belongs, the currency in which the project needs funding, deadline of the for donations, the goal amount of funding, launch time of the project, how much money pledged for the project before the deadline, the state of the project, the number of backers who supported the project, the country in which the project owners accommodate, the pledged amount converted into US dollars and the goal amount converted in US dollars. Therefore, we have fourteen columns, whose names are self-explanatory, for each project and an ID column to identify them. We will explain more about the structure of the dataset in the following data exploration section.

The techniques we used to process the data are Decision Tree and Naive Bayes Classifiers. Since our data is discretely structured, we decided to employ classification methods rather than a regression or another alternative. We believe that Decision Tree and Naive Bayes suite the nature of the dataset more. The details of the methods and their implementation are going to be clarified in the model section in the following pages. Then we will evaluate the results from the model and the success of the algorithms before concluding the paper.

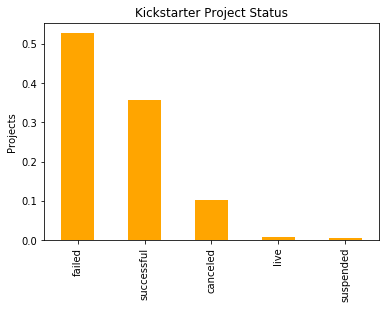
The main implications of our research are for modern economic theory within the creative destruction and entrepreneurship paradigms. The prominent economist Joseph Schumpeter was one of the first scholars who tackled with the concept of innovations in economy. In his book Capitalism, Socialism and Democracy (1942) he argued that “capitalism can only be understood as an evolutionary process of continuous innovation and 'creative destruction’”. In his own words Schumpeter (1942) explains the creative destruction phenomena as the “process of industrial mutation that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one”. The idea of creative destruction lies on the fact that the old and inefficient firms will be destroyed and put out of business environment by new and efficient firms to reach a better output/input ratio, i.e. more productive industries. In that sense innovation and R&D is the driving forces of economic growth and institutions, entrepreneurs and venture capitals are at the heart of the transition towards more productive economies. According to The Lisbon Council, a Brussels based think tank on economic and social policy making, as much as 85% of productivity growth in modern, developed economies is the direct result of innovation (van Ark, 2014). Our work stems from the theoretical and empirical importance of new businesses and start-ups. A mechanism that predicts which firms are worthy to be funded is extremely helpful in order to accelerate economic growth. Here by using the Kickstarter dataset, we rely on people’s judgements and intuitive senses to determine which projects have the quality for funding, thus we educate our algorithm on previous funding successes. We think that our work fills a gap within the interdisciplinary area of machine learning and economic theory.

**DATA EXPLORATION**

We start by redefining the data types in order to have the right structure for our analysis. Therefore, we first convert deadline and launched dates to datetime objects, then convert goal, pledged and backers columns to numeric objects. The next step is to rule out the unusable data from the dataset. We drop the all of the rows that have NaN values in them. This process eliminated 3801 observations and we are left with 374860 rows in the dataset.

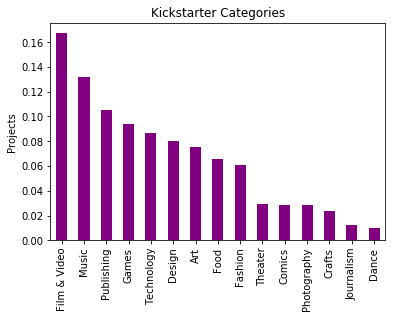
First, let us look at the distribution of project status. According to Figure 1, more than the half of the projects in our dataset are failed, which means that they could not collected the amount of money that was stated as the goal. On the other hand, nearly 35% of the projects could collect the required amount and thus labelled as successful. There are substantial number of projects, which are cancelled by the owner for various reasons. The live and suspended statuses are negligible.

*Figure 1: Project status*



We continue with examining the demand side of crowdfunding on Kickstarter. There are predefined categories for project proposals and every user chooses the most appropriate category to publish their idea. Figure 2 shows the distribution of the projects according the categories. The most popular categories are Film & Video, Music and Publishing, together they consist more than 40% of the projects. This situation is understandable when you consider the monopolistic structure of the entertainment and publication industries. There are a few numbers of firms offering finance and production opportunities in these markets and they do not invest money into the ideas that are not profitable. To break out of the rent seeking and profit-maximizing attitude of these industries, people are keen to propose their ideas for crowdfunding. The remaining categories can be clustered into two: Games, Technology, Design, Art, Food and Fashion consists 47% of the total and each of them is around 8% of the total; the remaining six categories, namely Theatre, Comics, Photography, Crafts, Journalism and Dance, are just 13% of the total projects.

*Figure 2: Projects' categories*

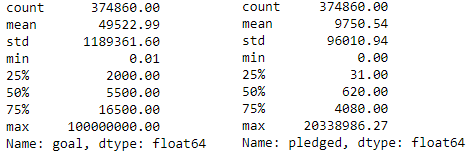


As the Figure 3 depicts more than three quarters of the projects are from United States, whereas nearly 10% are from Great Britain. The remaining twenty countries have marginally more than 10% share of the projects altogether. The distribution of the projects’ origin countries is heavily skewed which may lead into a possible sample selection bias. However, we cannot select another sample due to limitations of crowdfunding platforms. Our analysis would therefore be more accurate for US and the external validity for other countries might be limited.



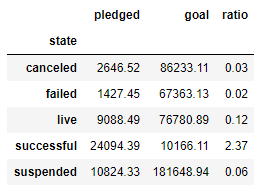
*Figure 3: Projects by origin country*

Now we will look at the descriptive statistics of goal and pledged variables. The mean value of the goal is around $50,000, but the standard deviation is extremely high around $1 million. There are some outliers in the data, since the median is $5,000. Pledged is more stable with mean around $10,000 and standard deviation of $96,000. The median is 620$. These numbers are noticeably lower across the board. We cannot get a good sense of project success by observing basic statistics. Therefore, we will get a sense for how successful each project generally is by creating a pledged/goal ratio on a per group basis.



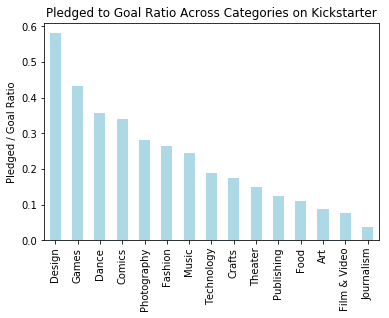
*Table 1: Descriptive statistics*

First, we create the pledged/goal ratio across state. Table 2 shows the grouped values of both variables and the ratio expressed according to states. The results are not surprising, since we were expecting successful projects to have a ratio which is bigger than one. On average, successful projects collected 2.37 times of aimed amount of money.



*Table 2: Pledged/Goal ratio by states of the projects*

Now, we are looking at the pledged to goal ratios across main categories of projects. The most salient category in terms of pledging success is Design. Projects in this category gathered on average 58% of the aimed amount, which is a respectable success. Followingly, Games, Dance and Comics projects are good enough to collect more than one third of the previously set goals. The three most popular category, which is Film & Video, is not very successful in money attraction and its pledged/goal ratio is slightly under 10%. This means on average Film & Video projects got one dollar for each ten dollar they wanted. Figure 4 below gives a clear picture of the ratios by categories.

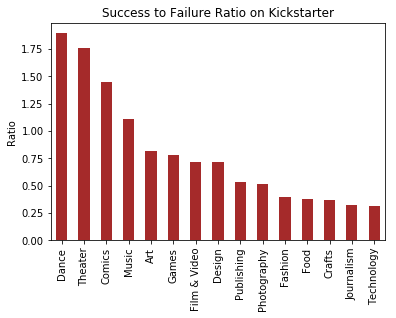


*Figure 4: Pledged/Goal ratio across categories*

To get a more macro feel for each category, we make the same analysis with successful to failed projects. The results are shown below in the Figure 5. Here we define three success levels:

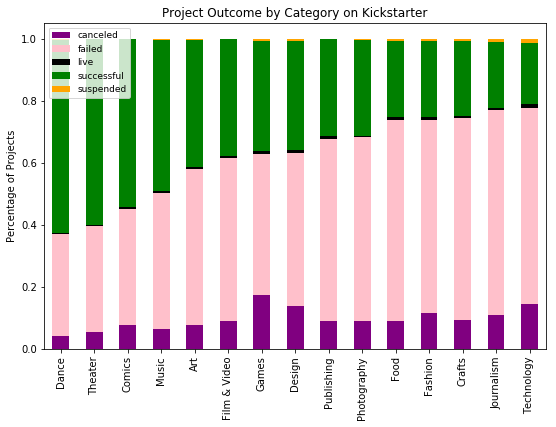
1. High Success: Dance, Theatre, Comics and Music projects are highly successful on average, so that they each have success ratio bigger than 1. In these categories there are more successful projects than there are failed ones.
2. Moderate Success: Following six categories, Art, Games, Film & Video, Design, Publishing and Photography, are somewhat successful on average. Their success ratio changes between 1 and .5, which means there are more failed projects than successful ones.
3. Shallow Success: We did not label the worst performing five categories as unsuccessful, because they each have some satisfactory number of successful projects in them. The lower bound of the success ratio is .31, but they are all below the half point.

Comparing the Figure 4 and Figure 5, the interesting cases are Theatre and Art on one hand and Design and Games on the other hand. The former duo has very low pledged to goal ratio, .15 and .09 respectively, whereas they are in high success category (Art is at the top of moderate success). The same discrepancy, but in the opposite way happened for the later duo. Design and Games projects were the most successful two categories in pledged to goal ratio, but they could get only moderate success. We could not explain the phenomena here other than the idiosyncratic preference differences of supporters for different categories.



*Figure 5: Success/Failure by categories*

Lastly, we would like to put everything together and get the overall picture in one graph. Figure 6 depicts the shares of the projects’ status across each category. In absolute terms the line-up does not change and the predefined success categories are still valid here.



*Figure 6: Project outcome by category*

**DATA CLEANING**

The dataset should be ready to implement machine learning methods; thus, we must arrange the variables according to our planned methods. The details of the ML methods will be clarified in the following section. Since we are going to use classification methods, we will first encode our categorical variables: category, main\_category and country. We have classified 158 categories, 14 main categories and 21 countries.

Next, we boil the outcome categories down. Since we are only interested in success, we will narrow down the dataset to only two outcomes, success and failure. In this way the analysis will be easier and the algorithms will run more smoothly.

Lastly, we will also split the data into three for the use towards different purposes. Train data will be 70% of the whole and the algorithm is going to be educated on this partition. Test and cross validation sets will each have 15% of the data and we will use them to assess the performance of each model.

**MODEL**

In our project, we applied two types of classifiers. The first one is Decision Tree Classifier and the second one is Naive Bayes Classifier.

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

A Naive Bayes classifier is a machine learning method which benefits from the probabilistic approach. The logic of the classifier is based on the Bayes theorem.

* P(A|B) =

It requires categorical variables to be applied. Its main algorithm is as follows:

* For a response with m classes C1,C2,...,Cm and the predictors x1, x2, ... ,xp, compute P(Ci|x1, . . . , xp)

In other words, Naive Bayes finds the probability of belonging to class C, given specified values of predictors. This method is convenient with big size data and depends on independence between predictive variables. In this regard; C, the target variable, is the variable of state with two values. Our predictive variables are 'category', 'main\_category', 'goal', and 'country'.

**RESULTS**

We reached 67% accuracy rate as designing model in the training part of the data at the decision tree method, which is good enough because after a point high accuracy rate implies a high error rate as well. As applying Naive Bayes model, the accuracy rate is around 44%, which is not predictive enough. However, the recall of the algorithm is as high as 99%, which points a very sensitive prediction performance. The confusion matrices are consistent with this comparison.

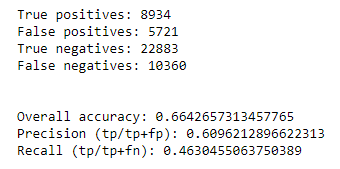


Table 3: Decision Tree Classifier results

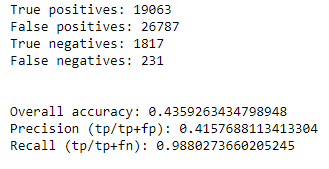


Table 4: Naive Bayes Classifier results

The accuracy rate indicates that this model can be improved but the more exploratory information about the data needs to be obtained. Most variables are not useful in a predictive manner. Also, there is a high correlation between variables except for the target one, which is a serious problem for any regression models.

For now, 67% accuracy rate is not sufficient but good enough to classify the dataset.

**CONCLUSION**

Our work’s foundations lie on the economic theory of creative destruction. Stemming from the questions we asked about the success of new ideas, we developed a framework to study the data from the crowdfunding website, Kickstarter. Our study puts light on the success patterns of projects. We divided the projects into three sets according to their state: i) high success ii) moderate success and iii) shallow success. Most successful categories are performative arts (Dance, Theatre and Music) and Comics. We believe that the determinants of success are associated with the preferences of supporters and also the nature of the crowdfunding process being public-benefit.

Our work puts the machine learning tools together with the economic theory. For this purpose, we developed two classification algorithms to predict the outcome of a project. The Decision Tree Classifier is accurate around 67%, which is an improvable overall accuracy. The Naive Bayes Classifier has worse overall accuracy, nearly 44%.

The performance of our algorithms can be improved, but there are problems along the way: More detailed and diverse data on crowdfunding outcomes is needed. This paper creates a link between the theory and advanced technological tools by practicing the newly emerging artificial intelligence algorithms on the economic theory. In this sense, we hope to pave the way for more research on this area and encourage interdisciplinary works in economics.

**REFERENCES**

Schumpeter, Joseph A. (1994). *Capitalism, Socialism and Democracy*. London: Routledge. pp. 81–84.

van Ark, Bart. (2014). “Productivity and Digitalisation in Europe: Paving the Road for Faster Growth”. *Lisbon Council Policy Brief Vol. 8, No. 1*.

Quinlan, R. (1983). "Learning efficient classification procedures". *Machine Learning: an artificial intelligence approach*, Michalski, Carbonell & Mitchell (eds.). p. 463-482.