HarvardX: PH125.9x Data Science

Successful Factors in Kickstarter Fundraising Projects

Hannah Tsang

28-3-2020

Contents

Introduction

Many startup companies may not have sufficient funds to develop and operate their business, so they may choose a crowdfunding platform to raise funds, one of the famous platforms is Kickstarter. Nevertheless, not all of them are successful to raise funds in Kickstarter, as a result, it is necessary to know which “types” of companies are more likely to be successful in fundraising with Kickstarter, so that a startup company which considers Kickstarter as the crowdfunding platform, will be more confident on whether it will be successful in the platform.

Aim of the project

In order to find out what companies are more likely to be successful in fund-raising with Kickstarter, we need to evaluate the characteristics of the successful companies as follows :

- The business nature and categories of the successful companies.
- The goal setting they expected for the amount of raised funds
- The place of origin of the successful companies.

Raw Dataset

```
URL      <- "https://drive.google.com/u/2/uc?id=1Yq-ShI0J_lPgF2rRFhJ9nP6kAJORYXcM&export=download"
destfile <- "C:/data_analysis/kickstarter.csv"
download.file(URL, destfile)
```

First, the Kickstarter Dataset is downloaded with the above code.

```
kickstarter <- read.csv(destfile)
summary(kickstarter)
```

Then, by running the above code in R Studio, a summary data will be shown, with the attributes as follows :

- ID
- Name
- Category
- Main Category
- Currency
- Deadline
- Goal
- Launch
- Pledged fund
- State
- Backers
- Country
- USD pledged
- USD_pledged_real
- USD_goal_real

Cleaned Dataset

Since we are only interested in the business nature, target amount of raised funds, and the business origin of the successful kickstarter project, we will clean some of the attribute data which are not relevant to our objectives. And the cleaned dataset will be as follows(here only the first 20 rows of dataset are shown as demonstration):

```
> kickstarter_cleaned <- kickstarter[,-c(1,2,3,5,6,8,9,11,13,15)]
> head(kickstarter_cleaned, 20)
   main_category   goal      state country usd_pledged_real
1     Publishing   1000     failed      GB             0.00
2   Film & Video  30000     failed      US          2421.00
3   Film & Video  45000     failed      US           220.00
4          Music   5000     failed      US             1.00
5   Film & Video  19500   canceled      US          1283.00
6           Food  50000 successful      US         52375.00
7           Food   1000 successful      US          1205.00
8           Food  25000     failed      US           453.00
9         Design 125000   canceled      US          8233.00
10  Film & Video  65000   canceled      US          6240.57
11    Publishing   2500     failed      CA             0.00
12         Music  12500 successful      US         12700.00
13        Crafts   5000     failed      US             0.00
14         Games 200000     failed      US             0.00
15         Games   5000 successful      GB        121857.33
16        Design   2500     failed      US           664.00
17        Comics   1500     failed      US           395.00
18    Publishing   3000     failed      US           789.00
19         Music    250 successful      US           250.00
20          Food   5000     failed      US          1781.00
```

As we can see above, after entering the blue-colored R codes, the remaining useful data attributes are “main_category”, “goal”, “state”, “country” and “usd_pledged_real”. “main_category” means the business nature of the kickstarter project ; “goal” means the target fund the kickstarter project owner(s) expected to raise ; “state” means whether the raised fund of the kickstarter project met the target, if the raised fund meets the target, the state will be shown as “successful, if not, the state will be shown as “failed” or “canceled” ; “country” means the origin place of the project ; and “usd_pledged_real” means the actual funding the kickstarter project had raised.

Data Analysis and discussions

Summary of Cleaned Data

```
> summary(kickstarter_cleaned)
```

After running the above blue-colored R code, the summarized dataset results will be shown as follows (here only the data attributes of “main_category”, “state” and “country” are shown as reference) :

```
   main_category
Film & Video: 63503
Music       : 51826
Publishing  : 39818
Games       : 35143
Technology  : 32485
Design      : 29970
(Other)     :125582
```

```
       state
failed    :197286
successful:133709
canceled  : 38717
undefined :  3555
live      :  2795
suspended :  1841
(Other)   :   424
```

```
     country
US      :292016
GB      : 33632
CA      : 14723
AU      :  7822
DE      :  4161
N,0"    :  3790
(Other): 22183
```

As we can see, the main categories of the kickstarter projects are Video film, Music, Publishing, Games, Technology and Design ; also, the main origin of the kickstarter projects are from the United States, and the rest of the main countries are United Kingdom, Canada, Australia and Germany, that means the majority of kickstarter projects are from the west. Besides, the successful rate among the 378327 kickstarter projects is around 35.3%, while the failed kickstarter projects is around 64.7%.

```
      goal                usd_pledged_real
 Min.   :        0   Min.   :       0
 1st Qu.:     2000   1st Qu.:      31
 Median :     5200   Median :     624
 Mean   :    49081   Mean   :    9059
 3rd Qu.:    16000   3rd Qu.:    4050
 Max.   :100000000   Max.   :20338986
```

Other than that, the mean goal set by all kickstarter fundraisers is $49081 USD, while the median is $5200 USD, which is much smaller than the mean value. It is probably because most of the kickstarter fundraisers tend to develop smaller projects which requires less money, so the median value is much smaller than the mean value ; besides, the reason why the mean value is much bigger than the median value is because the mean value is “amplified” by the big kickstarter projects which set goal with big amount of money, for example, the biggest amount of fundraising goal is up to 1,000,000,000, ie a billion USD, which indeed causes to mean value to be bigger even though the median is relatively small ($5200).

In contrast, the actual money donated by their supporters, ie the “usd_pledged_real”, has a mean value of $9059 USD and a median value of $624 USD, both of which are much smaller than those of the “goal” targeted by the fundraisers in Kickstarter. And also, the maximum amount of money ever donated to a single kickstarter project is $20,338,986 USD, which is around 20 millions USD, but still, it is much smaller than the biggest amount of “goal” set by a kickstarter fundraiser, which is around 1 billion.

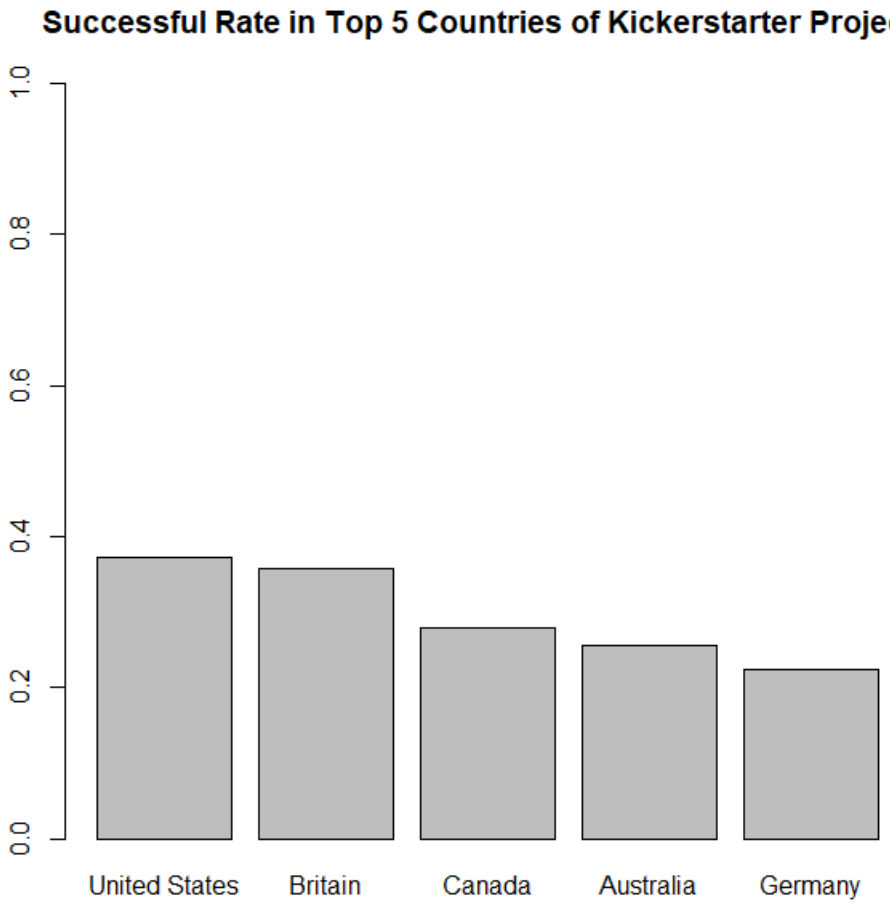
Therefore, it is revealed that the actual donated amount of money is smaller than that of goal expected by the fundraisers in Kickstarter.

But how does the probability of being successful or failed in kickstarter projects relate to the business nature, country origin or target amount of raised fund ?
As mentioned before in the Objective paragraph of this project , we would like to know 3 “factors” which bring the highest probability of success for launching a kickstarter project , ie Place of origin, Goal setting and the Business nature. Now let’s analyze the place of origin first.

Relationship between Successful Rate & Place of Business Origin

```
kickstarter <- read.csv("C:/data_analysis/kickstarter_data.csv")

gl <- kickstarter$goal

ste <- kickstarter$state

place <-kickstarter$country

count_success_us <- length(intersect(which(ste=="successful"), which(place=="US")))

count_fail_us <- length(intersect(which(ste!="successful"), which(place=="US")))

success_rate_us = (count_success_us / (count_success_us + count_fail_us))


count_success_gb <- length(intersect(which(ste=="successful"), which(place=="GB")))

count_fail_gb <- length(intersect(which(ste!="successful"), which(place=="GB")))

success_rate_gb = (count_success_gb / (count_success_gb + count_fail_gb))


count_success_ca <- length(intersect(which(ste=="successful"), which(place=="CA")))

count_fail_ca <- length(intersect(which(ste!="successful"), which(place=="CA")))

success_rate_ca = (count_success_ca / (count_success_ca + count_fail_ca))


count_success_au <- length(intersect(which(ste=="successful"), which(place=="AU")))

count_fail_au <- length(intersect(which(ste!="successful"), which(place=="AU")))

success_rate_au = (count_success_au / (count_success_au + count_fail_au))


count_success_de <- length(intersect(which(ste=="successful"), which(place=="DE")))

count_fail_de <- length(intersect(which(ste!="successful"), which(place=="DE")))

success_rate_de = (count_success_de / (count_success_de + count_fail_de))


combined <- c(success_rate_us, success_rate_gb, success_rate_ca, success_rate_au, success_rate_de)

barplot(combined, ylim=c(0,1), main="Successful Rate in Top 5 Countries of Kickerstarter Projects", horiz=F
        names.arg=c("United States", "Britain", "Canada", "Australia", "Germany"))

head(combined)
```

When we run the above code in R Studio, the following bar charts will be shown :



This bar-chart reveals that kickstarter projects in United States have the highest probability to be successful ; while those in Germany have the lowest chance to be successful ; and kickstarter projects in Britain, Canada and Australia are the 2nd, 3rd and 4th highest probability to be successful.

Relationship between Successful Rate & Goal expected by fundraisers

For better analysis, the goal of fund-raising expected by kickstarter project owners has been classified into 5 catagories as follows :

I : $5000 or below

II : $5001 to $15000

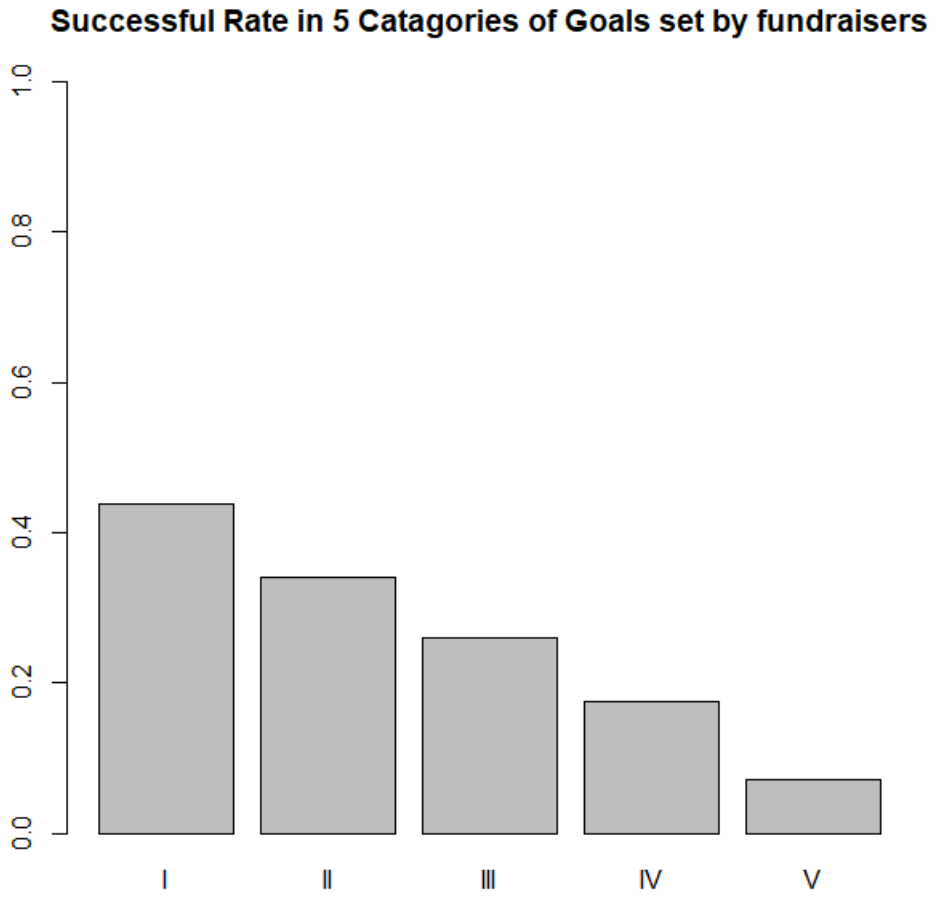
III : $15001 to $30000

IV : $30001 to $100000

V : $100001 or above

```
gl <- kickstarter$goal
count_success_I <- length(intersect(which(ste=="successful"), which(gl<=5000)))
count_fail_I <- length(intersect(which(ste!="successful"), which(gl<=5000)))
success_rate_I = (count_success_I / (count_success_I + count_fail_I))

count_success_II <- length(intersect(which(ste=="successful"), which(gl<=15000))) - count_success_I
count_fail_II <- length(intersect(which(ste!="successful"), which(gl<=15000))) - count_fail_I
success_rate_II = (count_success_II / (count_success_II + count_fail_II))

count_success_III <- length(intersect(which(ste=="successful"), which(gl<=30000))) - (count_success_I + count_success_II)
count_fail_III <- length(intersect(which(ste!="successful"), which(gl<=30000))) - (count_fail_I + count_fail_II)
success_rate_III = (count_success_III / (count_success_III + count_fail_III))

count_success_IV <- length(intersect(which(ste=="successful"), which(gl<=100000))) - (count_success_I + count_success_II + count_success_III)
count_fail_IV <- length(intersect(which(ste!="successful"), which(gl<=100000))) - (count_fail_I + count_fail_II + count_fail_III)
success_rate_IV = (count_success_IV / (count_success_IV + count_fail_IV))

count_success_V <- length(intersect(which(ste=="successful"), which(gl>100000)))
count_fail_V <- length(intersect(which(ste!="successful"), which(gl>100000)))
success_rate_V = (count_success_V / (count_success_V + count_fail_V))
combined <- c(success_rate_I, success_rate_II, success_rate_III, success_rate_IV, success_rate_V)

barplot(combined, ylim=c(0,1), main="Successful Rate in 5 Catagories of Goals set by fundraisers", horiz=FALSE,
        names.arg=c("I", "II", "III", "IV", "V"))
head(combined)
```

When we run the above code in R Studio, the following bar-chart will be shown



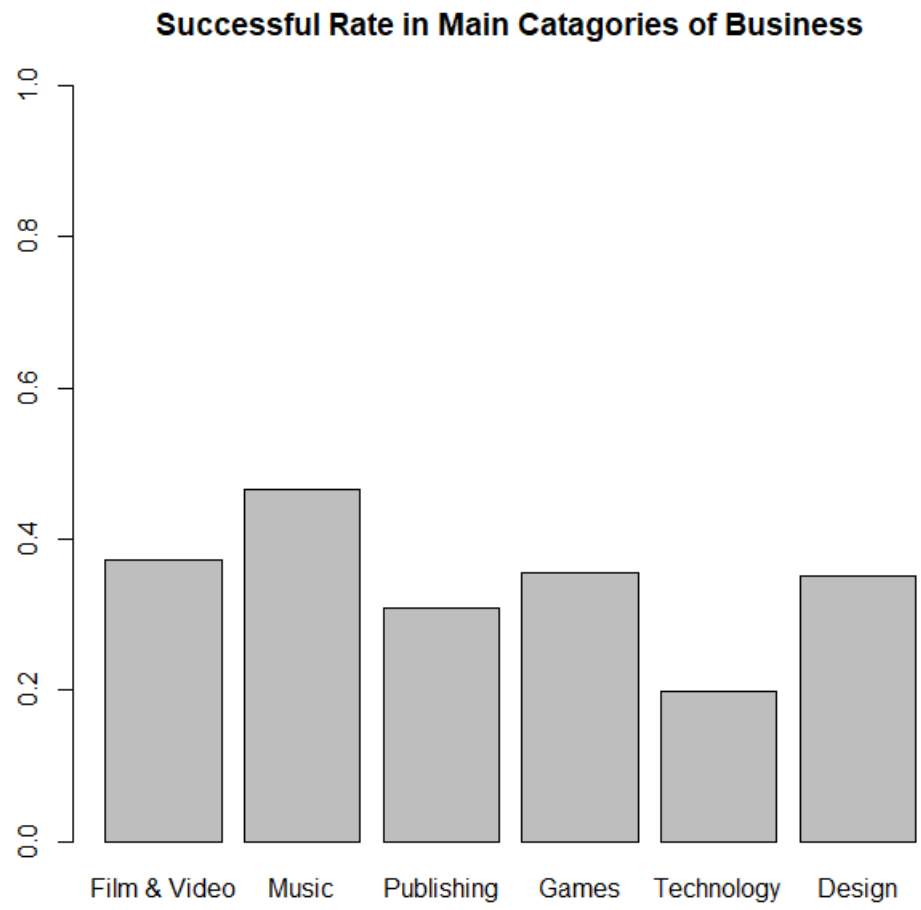
This chart reveals that the goal which belongs to Category I, ie below $5000, has the highest probability to be successful in kickstarter fundraising ; while the goal belonging to Category V, ie $100001 or above, has the lowest chance to be successful, ie meeting the goal expected by the fundraisers. The Category II, III and IV also has the 2nd, 3rd and 4th highest chance to be successful respectively.

Relationship between Successful Rate & the Business Nature

As shown in page 2-1, the major categories of business in kickstarter projects are Film & Video, Music, Publishing, Games, Technology and Design in descending order.

```
category <- kickstarter$main_category
count_success_film <- length(intersect(which(ste=="successful"), which(category=="Film & Video")))
count_fail_film <- length(intersect(which(ste!="successful"), which(category=="Film & Video")))
success_rate_film = (count_success_film / (count_success_film + count_fail_film))

count_success_music <- length(intersect(which(ste=="successful"), which(category=="Music")))
count_fail_music <- length(intersect(which(ste!="successful"), which(category=="Music")))
success_rate_music = (count_success_music / (count_success_music + count_fail_music))

count_success_publish <- length(intersect(which(ste=="successful"), which(category=="Publishing")))
count_fail_publish <- length(intersect(which(ste!="successful"), which(category=="Publishing")))
success_rate_publish = (count_success_publish / (count_success_publish + count_fail_publish))

count_success_game <- length(intersect(which(ste=="successful"), which(category=="Games")))
count_fail_game <- length(intersect(which(ste!="successful"), which(category=="Games")))
success_rate_game = (count_success_game / (count_success_game + count_fail_game))

count_success_tech <- length(intersect(which(ste=="successful"), which(category=="Technology")))
count_fail_tech <- length(intersect(which(ste!="successful"), which(category=="Technology")))
success_rate_tech = (count_success_tech / (count_success_tech + count_fail_tech))

count_success_design <- length(intersect(which(ste=="successful"), which(category=="Design")))
count_fail_design <- length(intersect(which(ste!="successful"), which(category=="Design")))
success_rate_design = (count_success_design / (count_success_design + count_fail_design))

combined <- c(success_rate_film, success_rate_music, success_rate_publish, success_rate_game, success_rate_tech, success_rate_design)

barplot(combined, ylim=c(0,1), main="Successful Rate in Main Catagories of Business", horiz=FALSE,
        names.arg=c("Film & Video", "Music","Publishing", "Games", "Technology", "Design"))
```

When we run the above code, the following bar-chart will be shown :



This bar-chart reveals that the kickstarter projects about Music have the highest probability to be successful, while those about Technology have the lowest chance to be successful.

Conclusion

According to the Data Analysis, along with dataset of Kickstarter and R codes with R Studio, we can conclude that Kickstarter projects from the United States, with fundraising goal below $5000, and with main category of Music as the project nature, will have the highest probability to be successful in fundraising with kickstarter platform.

Dataset Reference & Analysis Tools

Dataset : Kickstarter Project, by Mickael Mouille, downloaded from https://www.kaggle.com/kemical/kickstarter-projects

Analysis Tools : R Studio, downloaded from https://rstudio.com/