Kickstarter Data Project

Karan Goda | 460496371

Section 1: Problem

Background

Kickstarter is a global community that was founded in April 2009. Kickstarter allows people to invest in projects that they believe in, which are also created by other people who believe that their ideas will truly benefit the society. We can consider Kickstarter as a Crowdfunding organization that allows a group of people to back a project.

Since its launch, over 10 million people have backed a project(s) with over a total of $3.85 billion pledged with over 149,000 successfully funded projects. Some of the Kickstarter projects involved really famous people. The benefit on Kickstarter is that the creator maintains full control over their work, this allows creative projects to flourish and enable people to reinvent the wheel by investing their money in the products they support. Kickstarter has made it their mission to **help bring creative projects to life.**

From what we understand, Kickstarter contains a large number of projects on their platform which sprung from ideas from underdogs and start-ups. However the problem that Kickstarter now faces is that the failure funding rate of majority of these projects is high, not because the campaign creators had malicious intentions, but because the creators had a product that Kickstarter users did not want.

If the stakeholder Kickstarter was better equipped to predict which projects will fulfil funding, then the bar for future projects can be set much higher where Kickstarter will increase their customer's satisfaction rate and there will be an increase in the % of interesting projects that consumers will be interested in even after funding.

Research Questions

The project is built to understand the following research questions:

- Do certain categories of projects have a higher success funding rate than others?
- Does a longer project deadline mean greater chance of successful funding?
- Does a project funding fiat goal influence funding?
- With all the above questions answered, we can obtain information and predict which future projects have a higher chance of funding than others.

Section 2: Approach

The problem of Kickstarter being unable to effectively predict which projects will reaching funding will be solved by the following steps:

1. We will first study the data to eradicate and transform any missing or garbage data points in the rows by using Jupyter notebook to read the data, and then write relevant code to clean it.
2. Data analysis will be conducted on the resulting dataset after analysing, cleaning and transforming the dataset into something that is useful to this project to obtain the answers to our research questions.
3. Our last goal is to do data mining and training a classifier to predict which future Kickstarter projects have a higher success rate of receiving funding.

Our requirements for the results to ensure that the most useful possible analysis has been conducted is to be certain that all the research questions have been answered and the data is meaningful even after the project has been conducted for organizations.

Section 3: Data

Tools and techniques we will use

Our problem can be structured as a **multinomial logistic regression**.

The tools we can use to solve this are:

KNN Algorithm to find similar items, Random Forest, Boosting, Multinomial Logit Model, and Nested Model.

The tools we will use to clean and explore the data are:

1. **Jupyter Python** was used to clean and explore the dataset
2. **Unix** was considered but not used as Jupyter was effective in conducting all the cleaning and filtering processes we required
3. **SQL** will be considered for stage 2 of the project for further exploration of the data

Acquiring the data

The dataset was made available on **Kaggle** in 2018 from where it was obtained for further data analysis for this project. Kaggle gives all its users permission to modify, edit, and distribute the datasets present on its website as mentioned in their about page.

Originally there were 378661 rows and 15 columns in the dataset. This dataset was somewhat cleaned and processed by the original publishers which brings the question about the authenticity of the dataset as we do not know the data processing methods applied. We will however assume it is genuine and apply our processes. After our cleaning, transforming and polishing the dataset; we are left with 4451 rows and 10 columns in the dataset which we have called as **transformed_kickstarter_2017.csv** which is a 530 kb file.

Any missing value or type errors in the dataset were dealt with by:

1. If the Type error was where the value would be a number in word format, it would be converted to an integer.
2. If the value was null or garbage, then the entire tuple (row) would be removed from the dataset.

Describing the data

The data is a two dimensional csv file which is in a tabular format. See Appendix B[1.1] for the screenshot of our sample dataset.

See Appendix B[1.2] for the variable types of each of the columns in the dataset. The most interesting columns (attributes) in the dataset in terms of the problem we want to answer are:

- Main Category
- Deadline
- State
- USD Pledged
- USD Goal

The name and number of backers for the current analysis doesn't seem important. These categories however will be extremely important in helping us solve our research questions so that eventually we can predict with more accuracy which Kickstarter projects will be successful in funding.

Descriptive Stats

See Appendix B[1.3] for the descriptive stats table.

The mean is greater than Q2 for the attributes of our Kickstarter dataset. This suggests that the distribution is right skewed. There are not many useful insights from our descriptive stats table so we will instead group the Kickstarter projects by category and find the total pledged for each of these categories.

See Appendix B[1.4] for the grouped by category table.

From the plot we have created, we can see that the top 5 category funded Kickstarter projects that are funded are in design, games, technology, film and video, and art.

Bibliography

- https://www.kickstarter.com/about
- https://www.telegraph.co.uk/finance/businessclub/12073226/The-other-side-to-Kickstarter.html
- https://www.forbes.com/sites/suwcharmananderson/2012/11/30/kickstarter-dream-maker-or-promise-breaker/#47bd95e331aa
- https://www.kaggle.com/kemical/kickstarter-projects

Appendix B

[1.1] Sample Dataset:

launched	ID	name	main_category	currency	deadline	state	backers	country	usd_pledged_real	usd_goal_real
2017-01-01 00:35:20	1856478743	Podcast "El Valle de los Tercos": Latinos en S...	Journalism	USD	2017-02-06	successful	47	US	2028.00	1250.0
2017-01-01 00:52:32	1879437630	Resolute Magazine	Journalism	USD	2017-02-01	successful	16	US	502.00	500.0
2017-01-01 01:40:17	297142419	Pussycats: Sex, Drugs, & The Impossible #2	Comics	USD	2017-01-31	successful	59	US	1089.00	500.0
2017-01-01 01:54:38	1775305095	Octopeni Middle School: The Animated Series!	Film & Video	USD	2017-01-31	canceled	1	US	750.00	16000.0

Screenshot of a sample of our dataset

[1.2] Variable types of each of the columns in the dataset:

Variable	Type
Launched	Timestamp
ID	Numeric
Name	String
Main_category	String
Currency	String
Deadline	Timestamp
State	String
Backers	Numeric
Country	String
USD_pledged_real	Numeric
USD_goal_real	Numeric

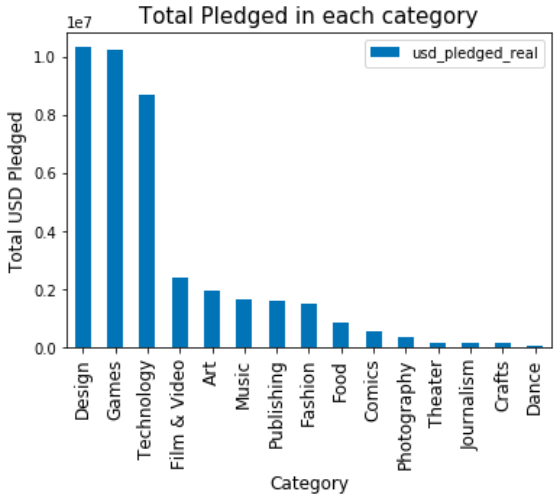
List of variables and their types

[1.3] Descriptive Stats:

	backers	usd_pledged_real	usd_goal_real
count	4451.0	4451.0	4451.0
mean	117.0	9163.0	56993.0
std	658.0	53096.0	1659753.0
min	0.0	0.0	0.0
25%	1.0	22.0	1481.0
50%	10.0	460.0	5000.0
75%	53.0	3268.0	15000.0
max	22834.0	1364835.0	107369867.0

The descriptive stats of the numeric columns of the dataset

[1.4] Grouped by category to get total USD pledged in each category table



Each different Kickstarter category with total pledged for that category