2019


# Predicting Success Rates of Kickstarter Projects

CAPSTONE PROJECT #1

TOM HALLOIN

# Table of Contents

## Problem Statement

Kickstarter is a crowdfunding website where users can raise funds to support creative projects. With enough funding, creators have the ability to launch new products, market new ideas, and possibly create new businesses. A successful project is one that passes its funding goal amount. If a project does not pass its goal amount, the project creators receive no funding. The goal is to determine the characteristics of a successful project and look for ways that creators can improve their chances of getting funding. To do this, I will create a statistical model to predict whether a Kickstarter project receives funding.

## The Basics of Kickstarter Projects

Here are the steps to create a basic project:

1.) Pick a category and subcategory
2.) Give project a title.
3.) Describe what you're creating (blurb).
4.) Pick your country from the following: Australia, Austria, Belgium, Canada, France, Denmark, Germany, Hong Kong, Ireland, Italy, Japan, Luxembourg, Mexico, New Zealand, Netherlands, Norway, Singapore, Spain, Sweden, Switzerland, United States, United Kingdom.
5.) Give the project an image
6.) Set a funding goal
7.) Set a campaign duration

The models will try to optimize these seven steps outside of the product image.

## The Dataset

There are many different Kickstarter data sets available, such as on Kaggle. I used the monthly JSON files on https://webrobots.io/kickstarter-datasets/ to get projects scraped from January 2016 to April 2019. For each project, I kept the following columns: backers_count, blurb, category name, category slug, country, currency, goal, launched_at, deadline, location_displayable_name, location_country, location_state, location_type, name, usd_pledged, slug, spotlight, staff_pick, static_usd_rate, and state.

## Data Wrangling

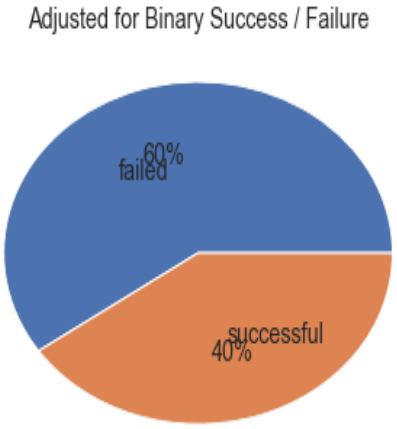
I created some new features: days before deadline, a goal converted from foreign currency to US dollars, the length of the blurb, and a binary final state. The original state had five different outcomes: success, failed, live, cancelled, and suspended. The binary final state converts these five outcomes to only success and failure by treating cancelled, suspended, and live projects that did not reach their funding goal as failures, while converting live projects that exceeded their funding goals as successful projects. This will be the target variable.

Five columns had missing values: blurb, location.displayable_name, location.country, location.state, and location.type. I replaced the missing blurbs with a '-', missing location.state values with '-', and missing location.type values with 'Miscellaneous'. There are instances where the country differed from the 'location.country'. The location.country matched the location.displayable_name (i.e. Jamaica to Kingston, Jamaica), so I kept that column and dropped the 'country' column.

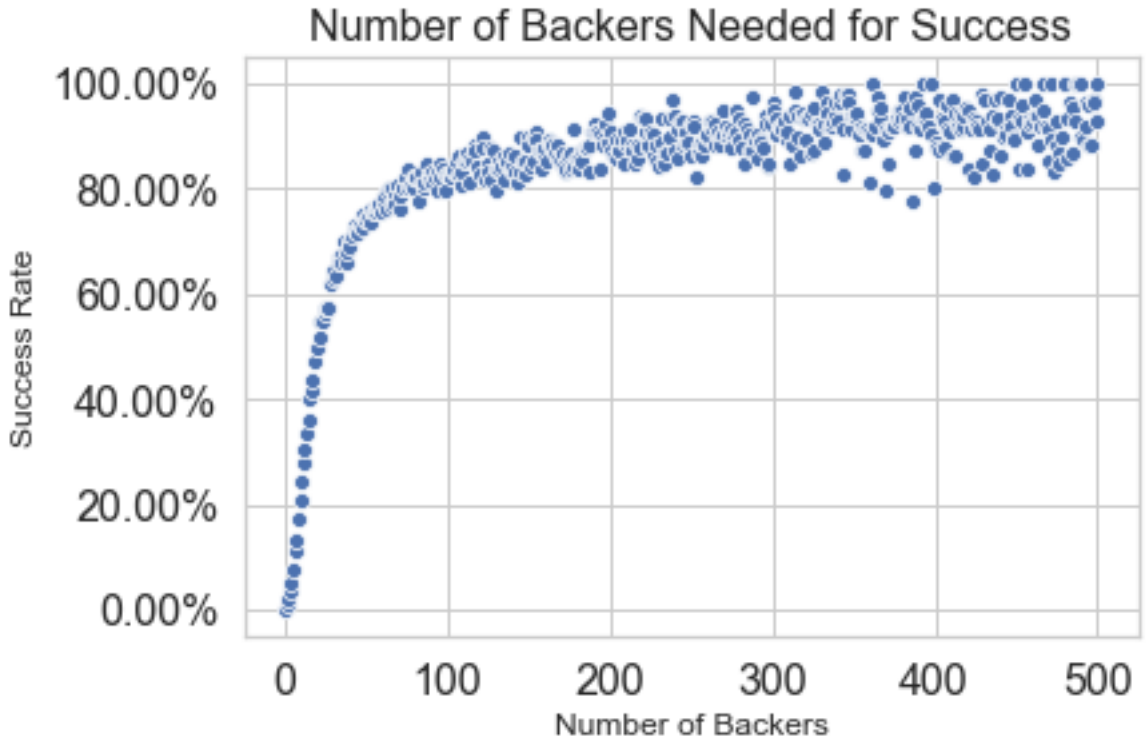
Most scrapes had duplicate projects, which I dropped. The final data frame has 430,714 projects and 29 variables.

## Exploratory Data Analysis

This is the original distribution of projects for each target category. When considering canceled, suspended, and live projects that have not met their goal as “failed”, there are more failed projects than successful ones.

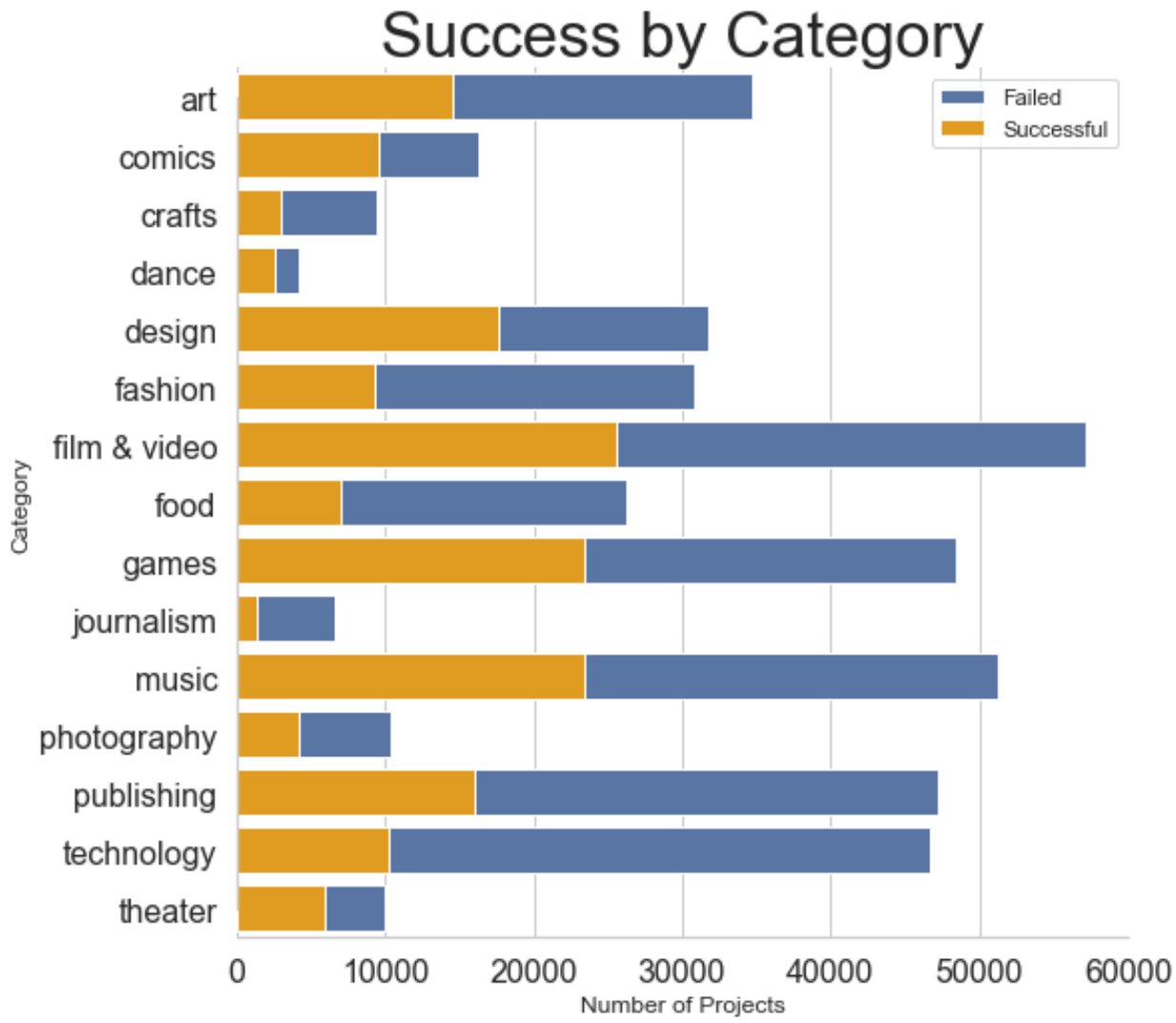


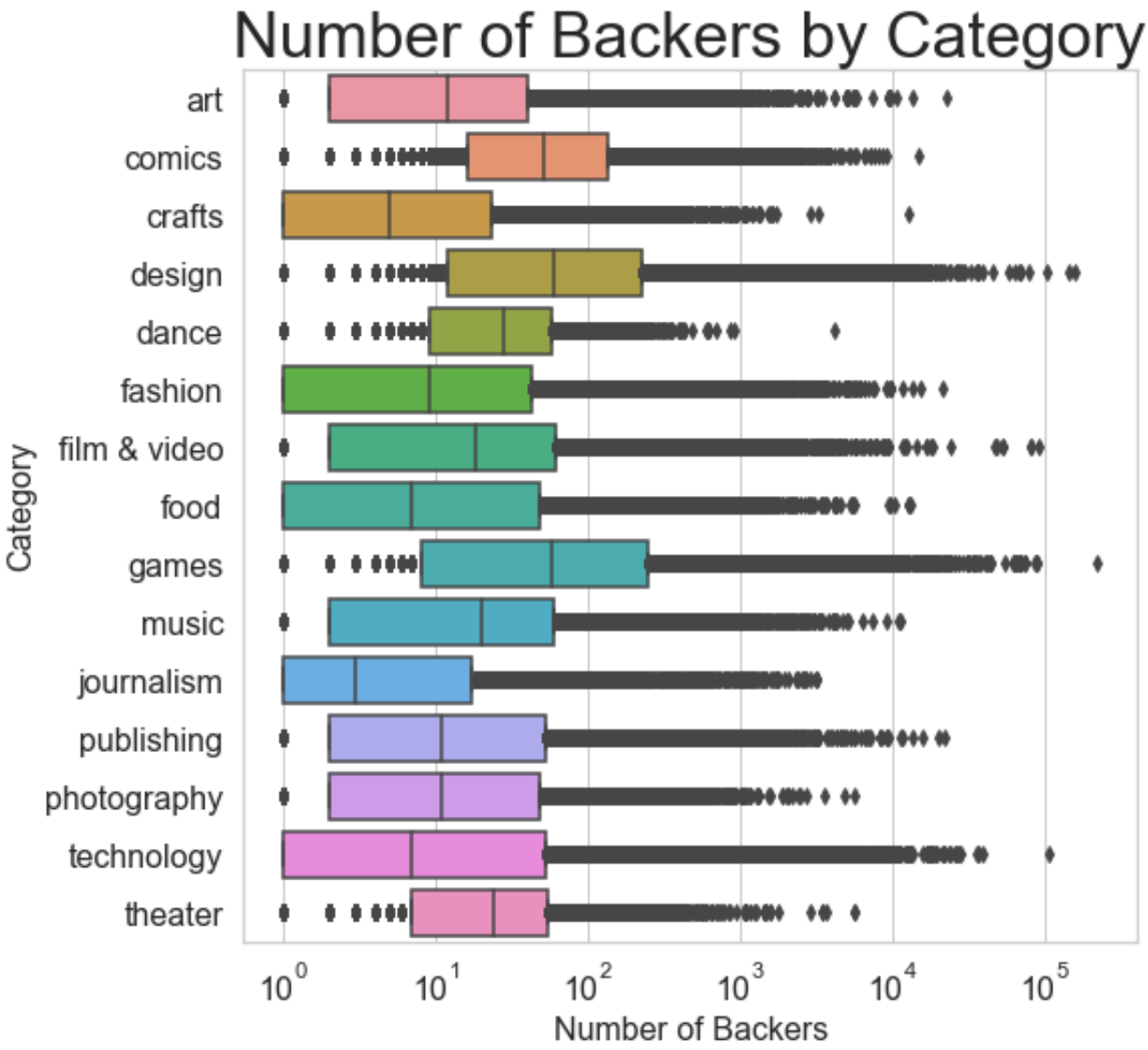

The most successful projects have the most people backing them. There is almost a direct correlation between the number of people backing a project and the project’s success rate. I will not include this in the final model, though. The goal is to build a model that will predict the success of a Kickstarter project before its launch. Models before launch have zero backers.



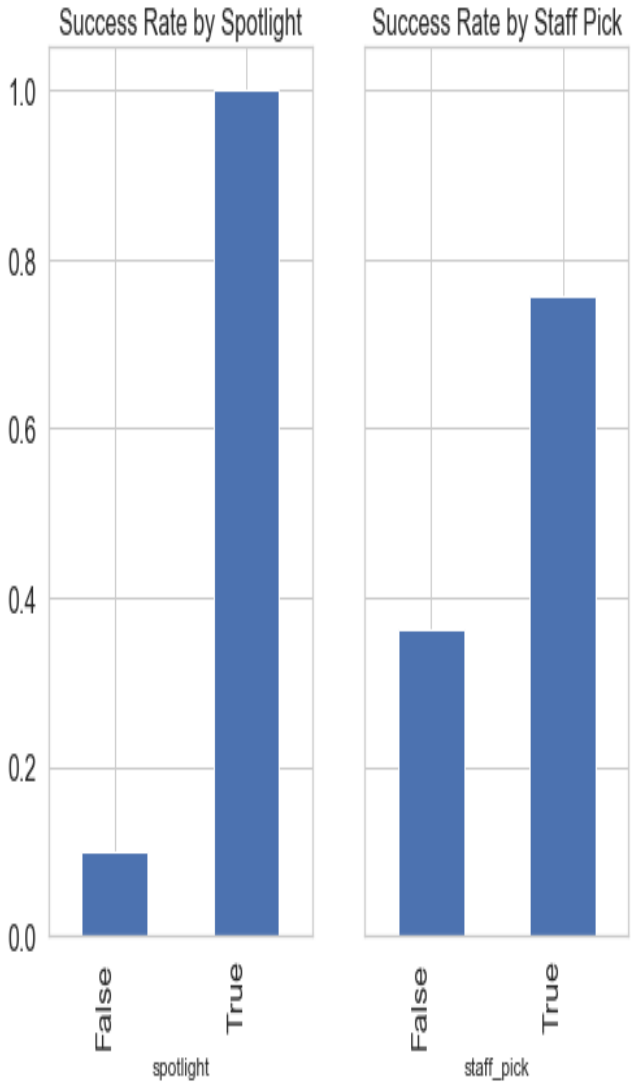

Kickstarter classifies their projects into 15 different categories: art, comics, crafts, design, film and video, food, games, music, journalism, publishing, photography, technology, theater, and

dance. Each one of these categories has a number of different sub-categories for a better description of the project. For the sake of EDA, I will stick with the broader categories. I included two additional charts by category: number of backers for each project and number of projects without any backers.

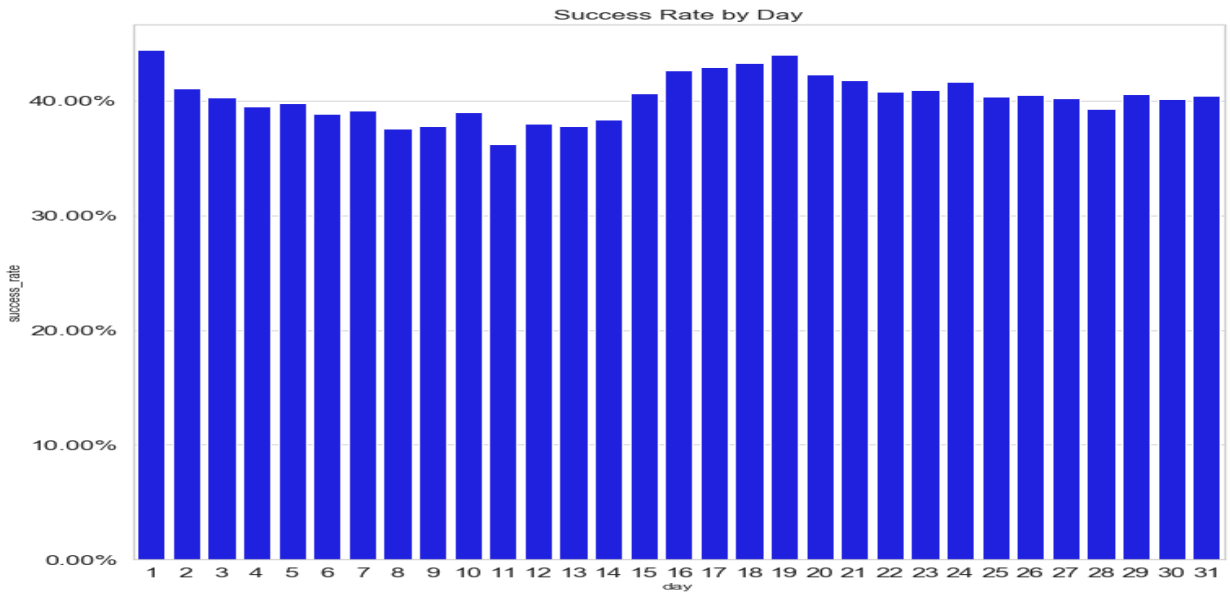


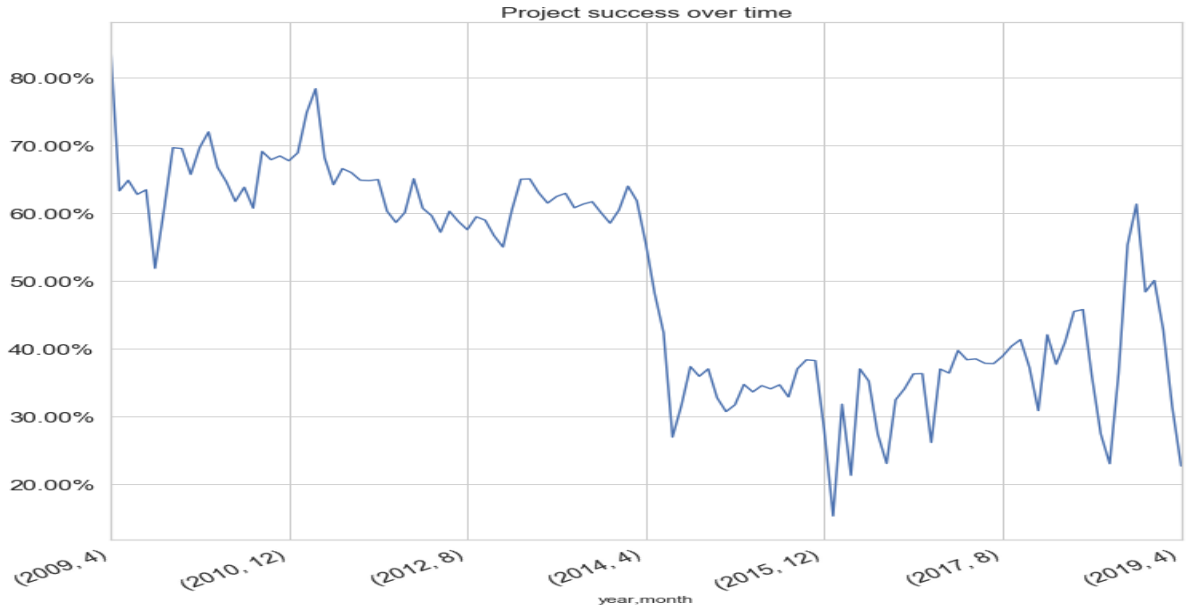

Number of Backers by Category
art
comics
crafts
design
dance
fashion
film & video
food
games
music
journalism
publishing
photography
technology
theater
Category
$10^0$
$10^1$
$10^2$
$10^3$
$10^4$
$10^5$
Number of Backers

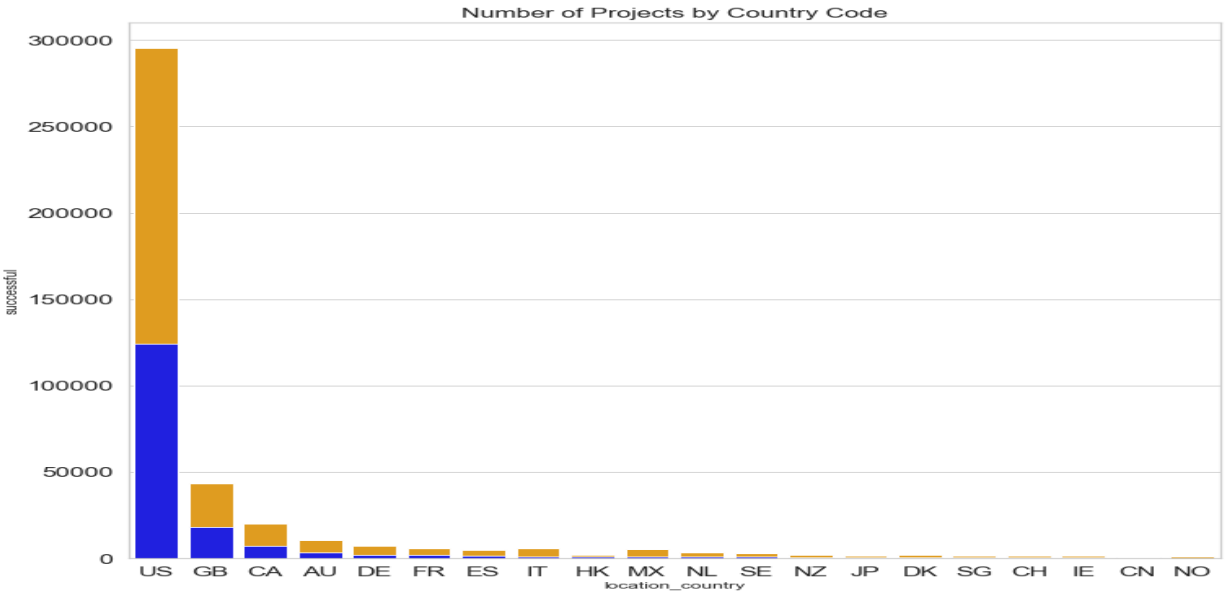
Kickstarter projects can have two additional designations: spotlight and staff_pick. The spotlight is for successful projects to share their story and progression. Staff picks are listed under “projects we love” and the first item that appears when searching for a project. It appears that putting a project in the spotlight or giving it a staff pick does increase the odds of the project receiving funding.

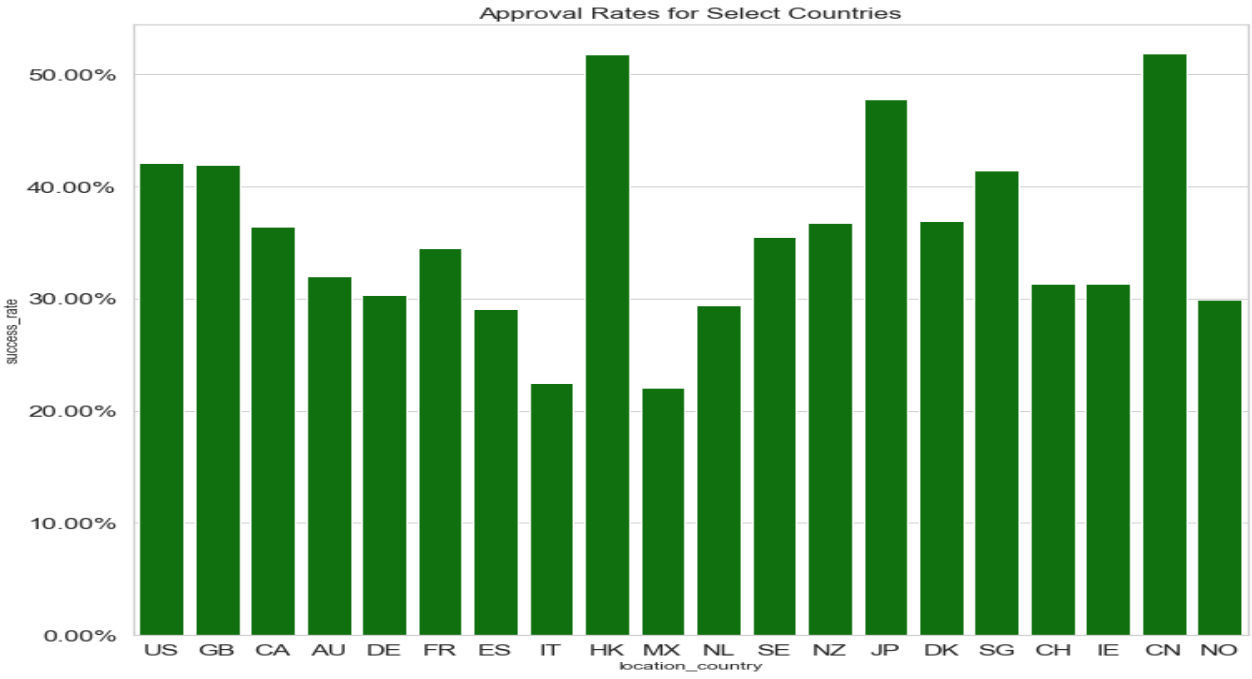



Over time, fewer Kickstarter projects have received funding. A part of this trend might come from the data collection and scraping process. The data come from JSONs scraped from January 2016 to April 2019. Each month contains scrapes from projects started before the scrape date. The monthly scrape in January 2016 contained 13 projects started in 2009. As Kickstarter became more popular, more people started to advertise their projects on the site. This led to more competition, driving the success rate down. It is also worth noting that some live projects are marked as “failures” even though their projects have not reached the funding deadline. That explains much of the drop in 2019 success rates.


Success Rate by Day
40.00%
30.00%
20.00%
10.00%
0.00%
success_rate
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
day


Project success over time
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
(2009, 4)
(2010, 12)
(2012, 8)
(2014, 4)
(2015, 12)
(2017, 8)
(2019, 4)
year,month

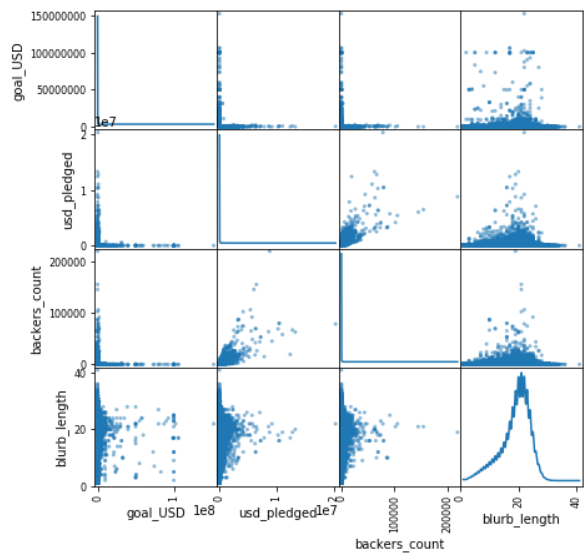


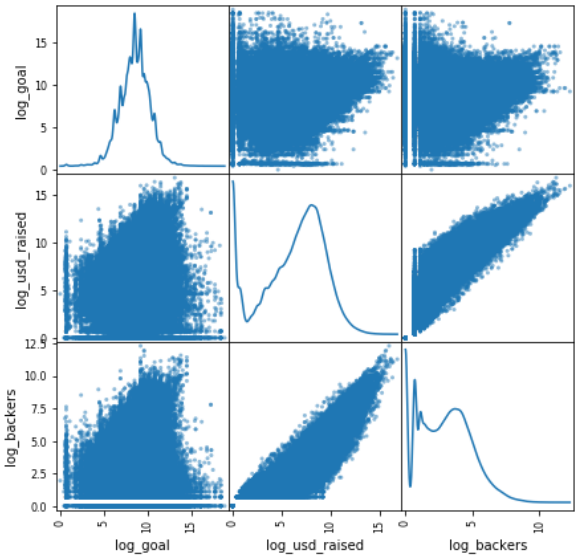

The United States has the most number of Kickstarter projects, followed by Great Britain, Canada, and Australia. Over 50 percent of the projects from Hong Kong, Switzerland, and Japan surpassed their funding goal.

## Statistical Inference

At this point, I have four quantitative variables: goal amount, amount pledged, number of backers, and blurb length. On the next page is a scatterplot for all four variables compared against each other. Ideally, all of the variables would be normally distributed, but that is not the case here. Some projects have goal amounts over $10,000,000, while most have much less than that. It also appears that the number of backers correlates to the amount raised. Before log transformation, I want to show a correlation matrix of these four variables. While there isn't an association between the goal amount and the actual amount raised, there is a positive association between the number of backers a project has and the amount raised.




*Pre Log Transformation*




*Post log transformation*

| | goal_USD | usd_pledged | backers_count | blurb_length |
|---|---|---|---|---|
| goal_USD | 1 | 0.0076 | 0.0057 | -0.0036 |
| usd_pledged | 0.0076 | 1 | 0.77 | -0.0057 |
| backers_count | 0.0057 | 0.77 | 1 | -0.0078 |
| blurb_length | -0.0036 | -0.0057 | -0.0078 | 1 |

| | log_goal | log_usd_raised | log_backers | blurb_length |
|---|---|---|---|---|
| log_goal | 1 | 0.13 | 0.11 | 0.026 |
| log_usd_raised | 0.13 | 1 | 0.93 | -0.026 |
| log_backers | 0.11 | 0.93 | 1 | -0.035 |
| blurb_length | 0.026 | -0.026 | -0.035 | 1 |

For the categorical variables, I performed a series of Chi-Square tests to determine whether there was any statistically significant difference between the expected frequency and actual frequency of project approval rates. Below is a table showing the null and alternative hypothesis, the test value result, and conclusion. All of the p-values turned out to be small enough so that they were considered to be zero after significant rounding error, so we can assume all of these variables matter in terms of predicting Kickstarter project success.

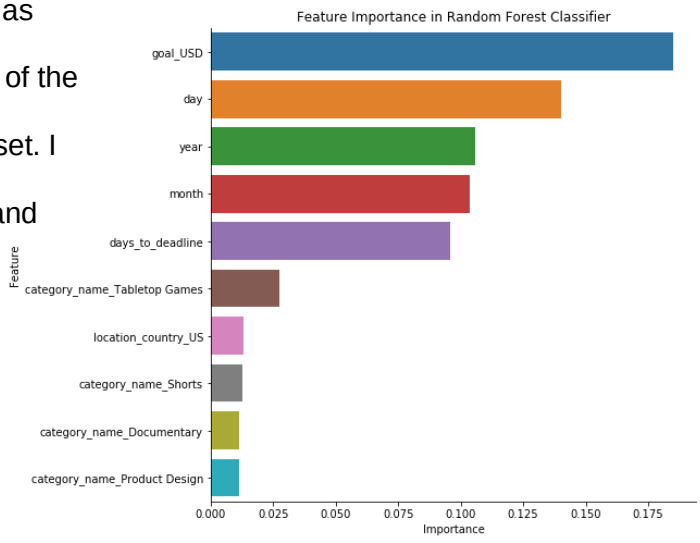
| Null Hypothesis | Statistic | Critical Value | P-Value | Conclusion |
|---|---|---|---|---|
| Project success does not depend on the subcategory of the project. | 70763.73 | 188.332 | 0 | Reject |
| Project success does not depend on whether a project is a staff pick. | 20262.06 | 3.841 | 0 | Reject |
| Project success does not depend on the category of the project. | 17017.87 | 23.685 | 0 | Reject |
| It does not matter what country a project is from. | 8651.73 | 23.685 | 0 | Reject |
| Project success does not depend on the day of when the project was released. | 984.8 | 15.507 | 0 | Reject |
| Project success does not depend on the month of when the project was released. | 597.53 | 19.675 | 0 | Reject |

## Machine Learning

I used five different machine learning algorithms: logistic regression, naïve bayes, random forest, stochastic gradient descent, and XGBoost. Logistic regression is a common algorithm to use for classification projects such as this one. Naïve Bayes is useful in particular for text analysis. Besides classification, I can use Naïve Bayes to find the probability of a

successful project given the word is in the title or description. This is useful to find popular topic ideas. Random forests are a series of decision trees and work well with a series of weak predictors such as the features available in this model. Both Stochastic Gradient Descent and XGBoost models have more hyperparameters to tune in order to find an optimal model. Stochastic Gradient Descent is useful given the size and nature of this dataset. The version of XGBoost I will be using is a random forest classifier.

I used the following features for my model: the goal amount, project release date, the number of days until the deadline, the category name, the subcategory name, and the project's location. I saved project titles and descriptions for another model that will use just the text data. I split the day into month, day, and year to use as independent features. I randomly assigned 2/3 of the data for the training set and 1/3 for the testing set. I scored each model by classification accuracy and provide precision, recall, and F1 scores. For all models except Naïve Bayes, I used a randomized grid search to fine tune hyperparameters. This improved the final accuracy scores to an extent, but not much.




The best model was a Random Forest classifier, with an accuracy score of 95% on the training data and 75% accuracy on the testing data. As evidenced by the chart, the most important features were the goal amount, followed by the day of the month, the year, the month, and the number of days until the deadline. Other important features included whether the project was a tabletop game or whether it launched in the United States. This model does not try to interpret direction. For instance, it does not indicate that projects with a higher goal amount have a lower success rate.

To find successful project ideas, I fit a Naïve Bayes model using just the text data from the title and description. The output of this model is the conditional probability of a project's success given the word's appearance in the title or description. After performing some data cleaning using Natural Language Processing, the words with the highest probability were 'miniature(s)', 'printable', 'chtulhu', 'pathfinder', 'titanium', 'edc', 'terrain', 'dice', and 'stretch'. *Sandy Petersen's Cthulhu Mythos for Pathfinder, for example,* sells a series of printable miniatures playable for a spinoff of *Dungeons and Dragons*. The project raised over $200,000 from over 2,000 backers. "EDC" stands for every-day carry – in other words, stuff people carry in their pockets. Many of these have titanium in them, such as pocket knives, watches, and keychains.


*Sandy Petersen's successful Cthulhu Mythos for Pathfinder project raised over $200,000. See project here.*

Common words in projects that did not reach funding goals include 'app', 'application', 'cater', 'network', and 'website'. All of these types of projects have alternative methods of support. The best way to support an app company is to buy their app. The best way to support a catering company is to buy their food. The best way to support a website is to visit the website. A backer does not need to use Kickstarter to support any of these projects. It is also possible that many of these markets are highly saturated with free alternatives.

## Recommendations based on analysis

- The more backers a project has, the more likely the project will be successful.
- The smaller the goal, the more likely the project will be successful.
- Plan to release a project on the 1st of the month or right after the 15th.

- There have been many successful projects based off of tabletop games, particularly those related to *Dungeons and Dragons.*
- Other successful projects include “every-day carry” items.
- Apps and websites struggle to get funding.
- Shorter campaign durations are better.

## Conclusion

This concludes a walkthrough of the data science process. It started with data. From there, I tried to find patterns. From these patterns, I formed a hypothesis that I could predict the likelihood of a Kickstarter project receiving funding using machine learning algorithms. Based on the final model accuracy score of 75%, I created a model proving the hypothesis. In addition, text analysis recommended project ideas for future backers.

Special thanks to Liang Kuang for his assistance in completing this report.