Kickstarter Project: Predicting Success

Final Report

Kickstarter

Crowdfunding Platform

$4,000,000,000+ towards projects

168,000+ successfully funded projects

16,000,000+ users

55,000,000+ total pledges

The Problem

Only 37.22% of Kickstarters have been successful

Creators may have to abandon their first go-round with a Kickstarter campaign

Proposed Solution

Given what we know about projects such as:

Description Lengths, Goal Amount, Category, Country of Origin, Time of Creation, Length of Campaign

We would like to know the success rate of whether or not a project will succeed.

To do this, we’d like a model that can take various inputs to answer whether a project will reach its goal.

Kickstarter Dataset

Kickstarter does not provide a public API

Web Robots (webrobots.io) scraped Kickstarter to provide this data

- 2019-05-16
- 210,000 rows with 37 columns
- 56 csv files

Additional web scraping utilizing Google Cloud Platform for more information

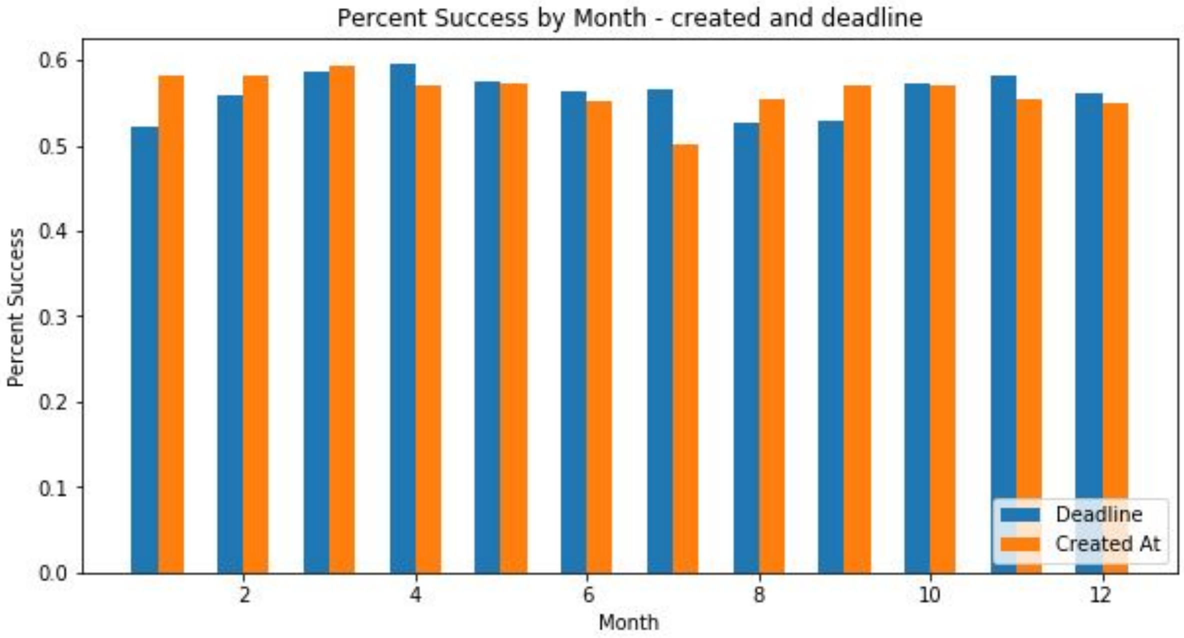
- Use URLs from the WebRobots data
- Pull information such as descriptions & rewards

Exploratory Analysis

Projects by Month

Projects started in March are the most successful while projects started in July are the least.

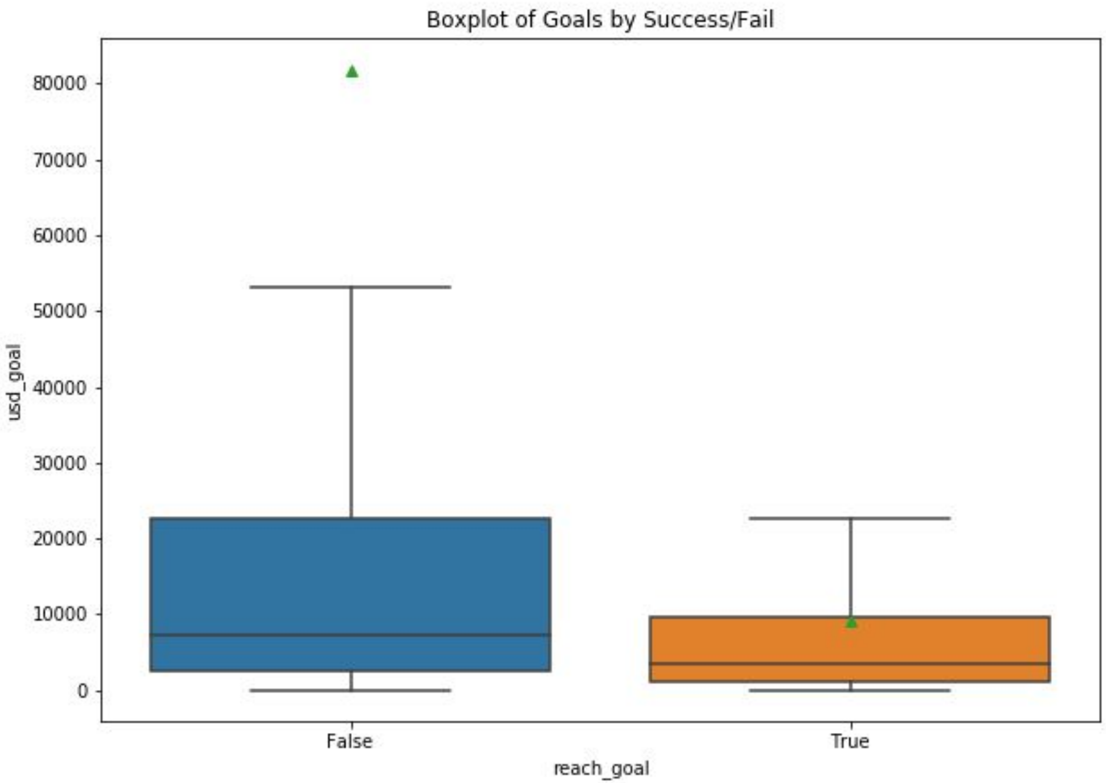
Alternatively, projects finishing in January are the least successful while finishing in April has the highest success.



Success to Goal (USD)

In general, projects which are successfully funded have much lower mean goals (USD) than projects who have failed.

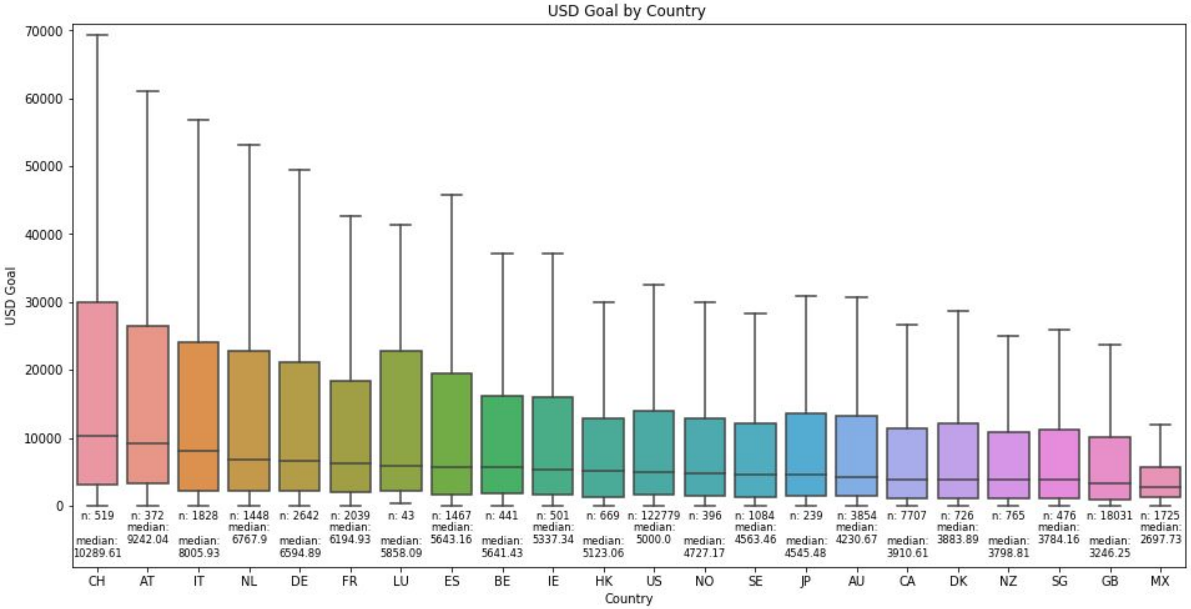
Additionally, the median of successful projects are also lower than the median of failed projects.



Country of Origin

In the boxplot of Goals (USD) by Country of Origin, we can see China has the highest median goal at $10,289. The next would be Austria and Italy at $9,242 and $8,005, respectively.

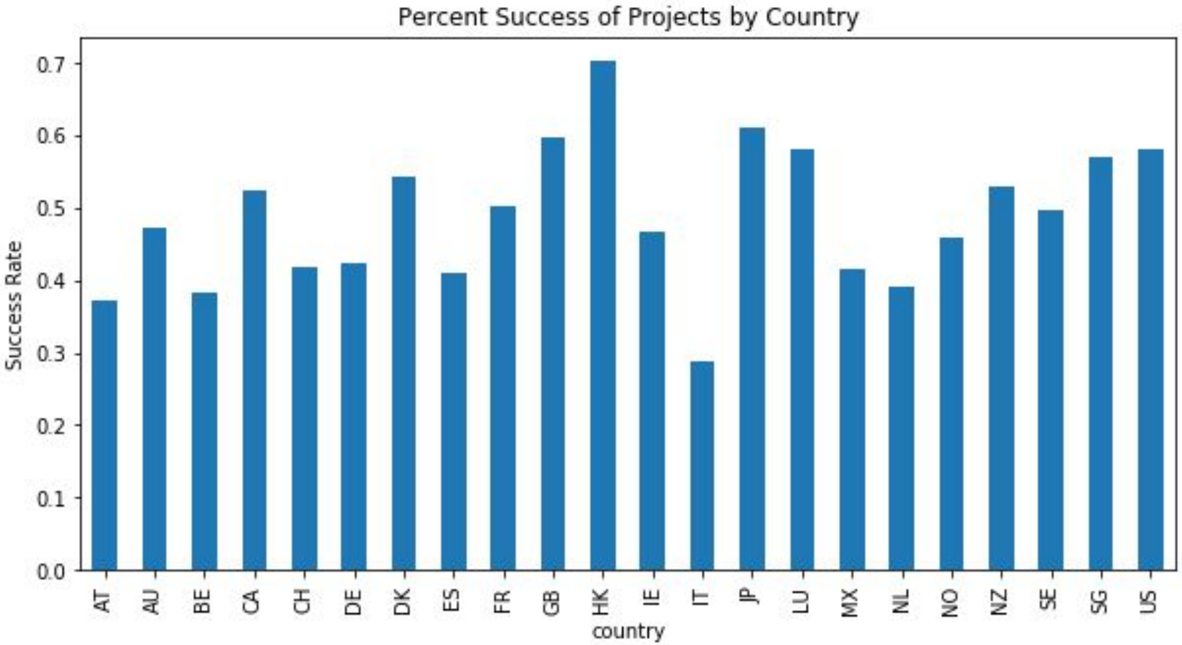
After, median goals seem to balance off around 5,000 to 2,000 USD. The spread also looks to decrease at median goals decrease.



Country of Origin

However, the three countries who have the highest median goals do not have the highest success rates. Italy actually has the lowest success rate for projects.

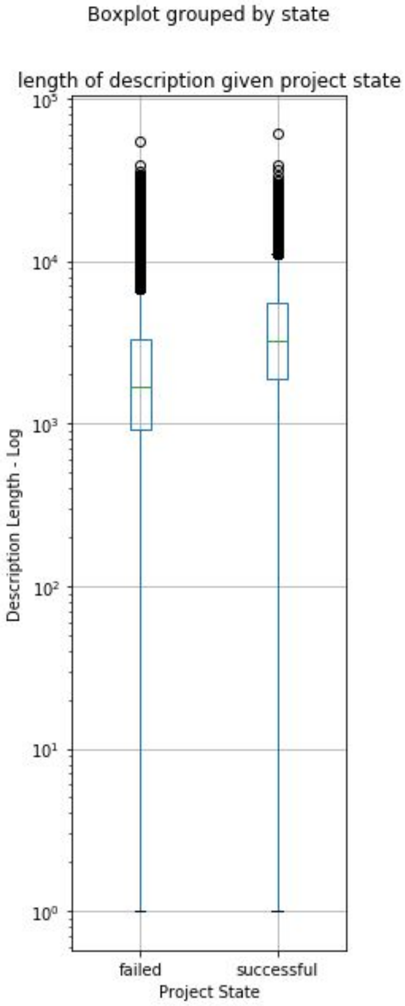
Hong Kong and Japan, countries with a median goal of 5,100 and 4,500 have the two highest success rates but are also on the lower end of the amount of projects from the country.



Project Descriptions

In the boxplot for description lengths of successful and failed projects, we can see that the median description length of successful projects is higher than failed projects.

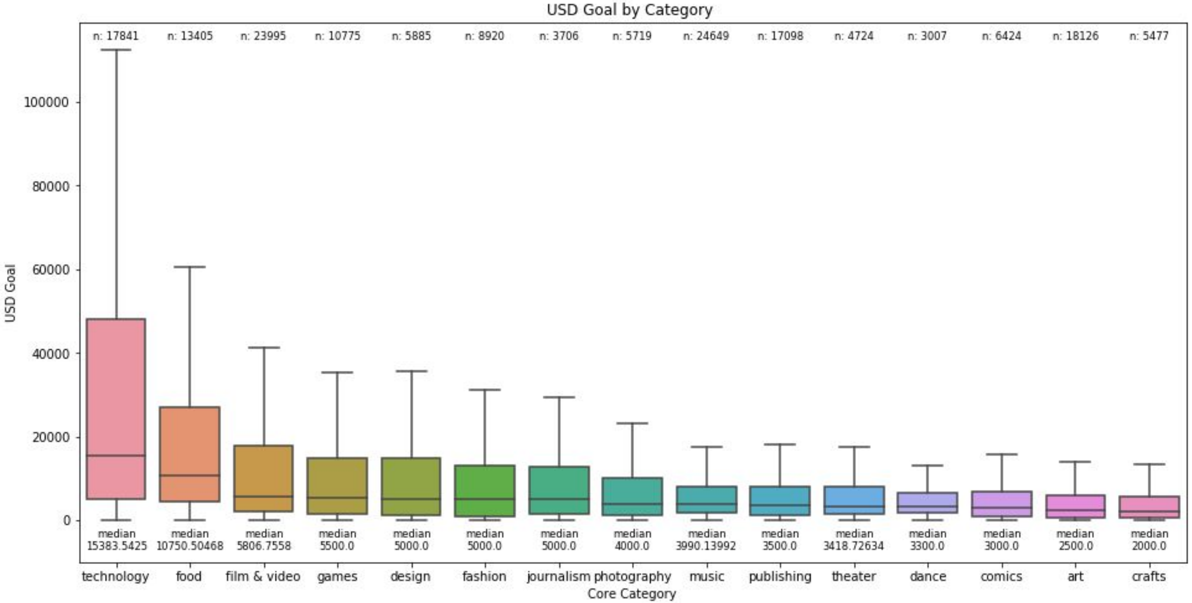
It's worth noting that some projects, both successful and failed, have descriptions of length 0, meaning they have used images and/or video to describe their project.



Category

Looking at categories to their Goal (USD), we can see that technology has the highest median goal at $15,383. Followed by food, then film & video.

Size of each category varies widely ranging from 24,000 to 3,000 in each category.

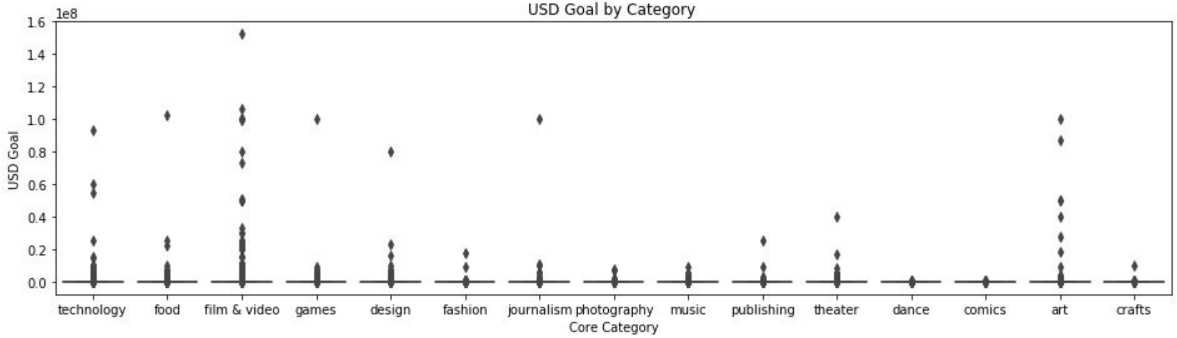


Category

Looking at the outliers for the previous boxplot, we can see that many categories are afflicted by high outliers such as technology, film & video, and even art.

Film & Video has the highest outlier goal at 1.6 million.

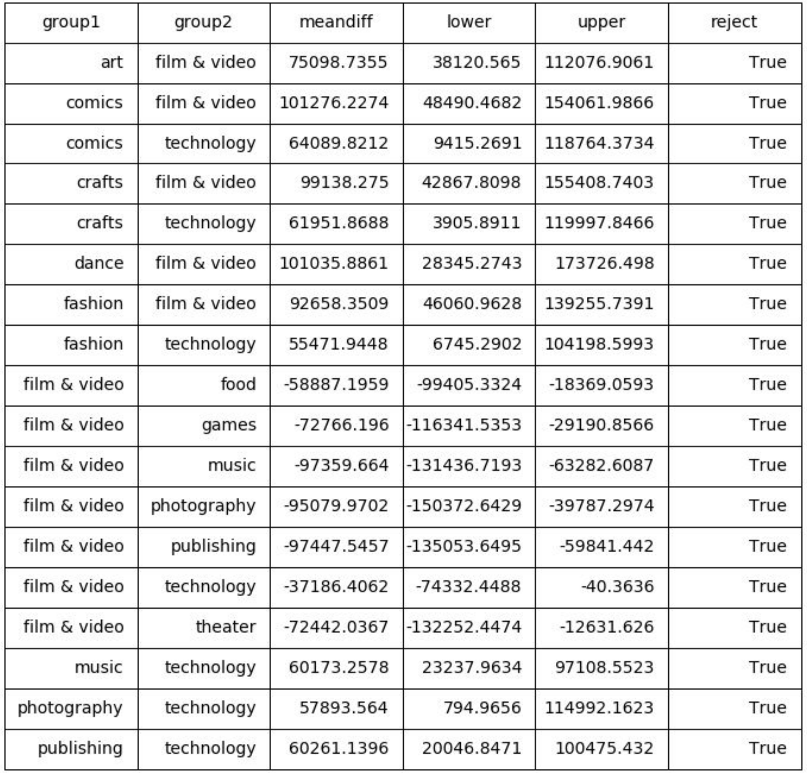
Tech, Film & Video, and Art have outliers across their range but some categories like Journalism are afflicted with gaps of goals.



Category

Performing a Tukey test, we can see that technology and film & video are the most different from the other categories when it comes to their goals.

Intuitively, this makes sense since tech and videography projects would seem to expend more money than other projects.



group1	group2	meandiff	lower	upper	reject
art	film & video	75098.7355	38120.565	112076.9061	True
comics	film & video	101276.2274	48490.4682	154061.9866	True
comics	technology	64089.8212	9415.2691	118764.3734	True
crafts	film & video	99138.275	42867.8098	155408.7403	True
crafts	technology	61951.8688	3905.8911	119997.8466	True
dance	film & video	101035.8861	28345.2743	173726.498	True
fashion	film & video	92658.3509	46060.9628	139255.7391	True
fashion	technology	55471.9448	6745.2902	104198.5993	True
film & video	food	-58887.1959	-99405.3324	-18369.0593	True
film & video	games	-72766.196	-116341.5353	-29190.8566	True
film & video	music	-97359.664	-131436.7193	-63282.6087	True
film & video	photography	-95079.9702	-150372.6429	-39787.2974	True
film & video	publishing	-97447.5457	-135053.6495	-59841.442	True
film & video	technology	-37186.4062	-74332.4488	-40.3636	True
film & video	theater	-72442.0367	-132252.4474	-12631.626	True
music	technology	60173.2578	23237.9634	97108.5523	True
photography	technology	57893.564	794.9656	114992.1623	True
publishing	technology	60261.1396	20046.8471	100475.432	True

Category + Description

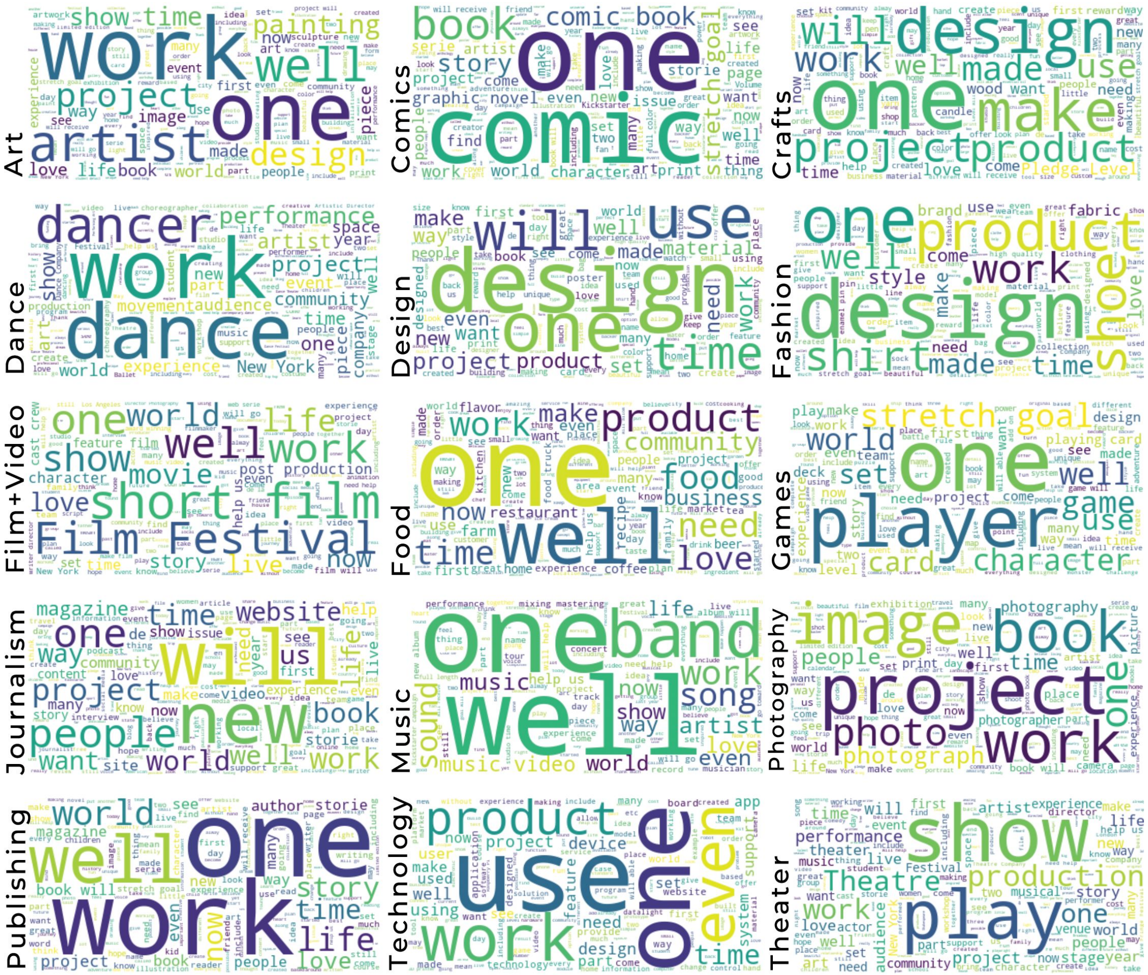
We can also explore how each category has certain word choices for their project.

One for “one of a kind” or “one perfect brew”

Design relates to categories like design, fashion, and art.

Device is seen in technology

Recipe & **Farm** appear in food

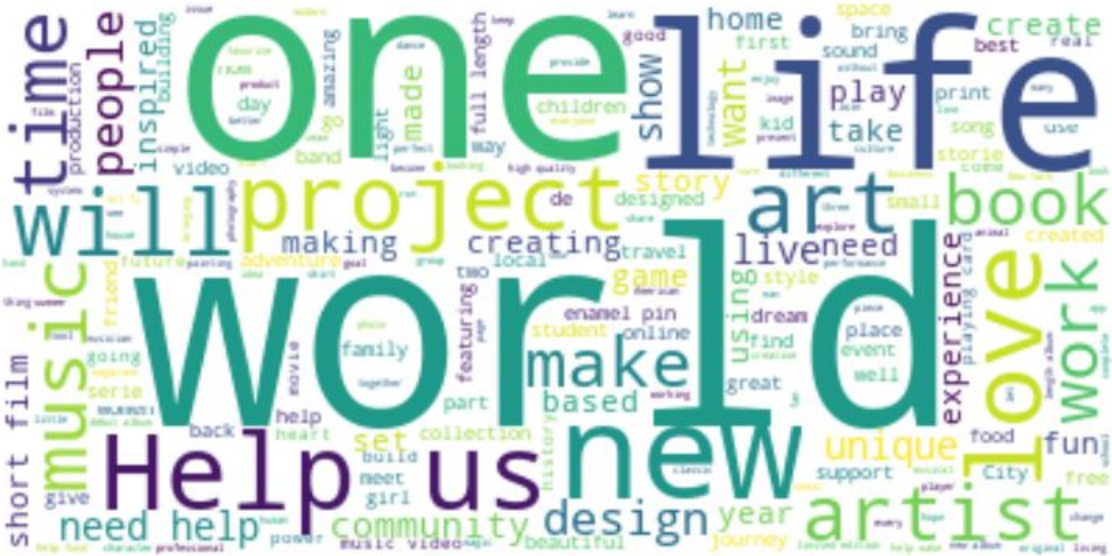


Blurbs

Across all projects, we can see what word choices they use to entice prospective backers to click their campaign.

One is used as superlatives i.e. “One of the most”

World is used in different manners: the world as a whole and world of their projects
- “world premier”, “around the world”
- “world of wine”, “art world”



Blurbs

We can also take a look at predictive features of words using MultinomialNB.

Few have >.90 as their predictors

Unusual ones such as “EDC” for Design and “28mm” for Games.

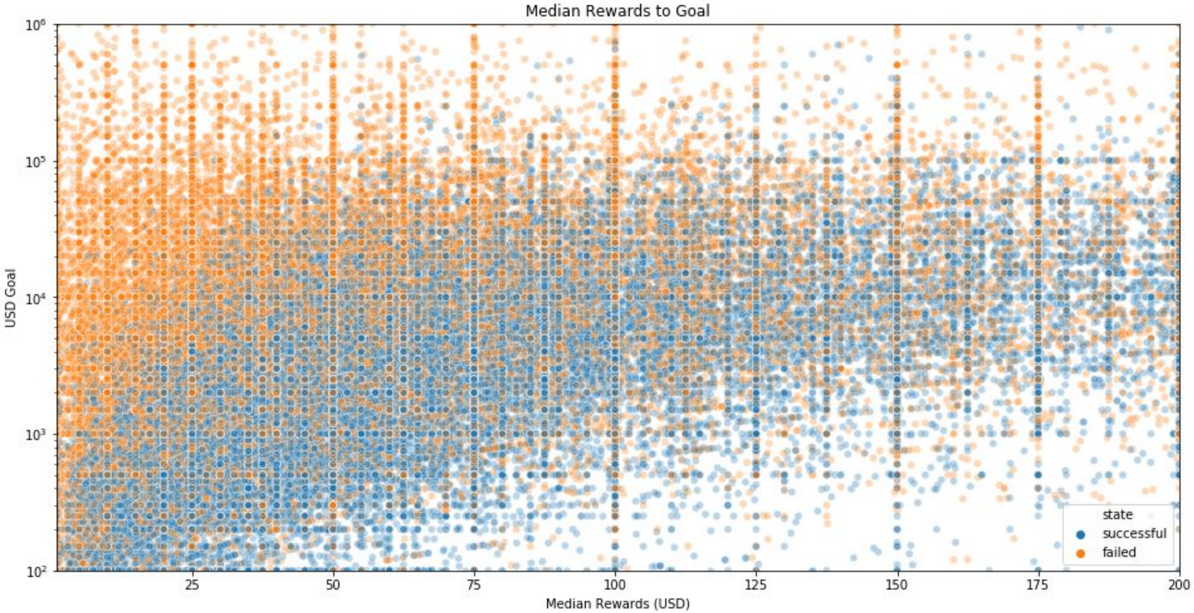
International keywords appear in Film & Video as “cortometraje” and in Theater as “Teatro”

ART: enamel | 0.93 pin | 0.92 sketchbook | 0.92 coloring | 0.91 78 | 0.90	COMICS: issue | 0.96 collected | 0.95 collection | 0.93 high | 0.93 face | 0.93	CRAFTS: plush | 0.91 enamel | 0.88 timeless | 0.84 giant | 0.82 orphan | 0.82
DANCE: work | 0.96 collaboration | 0.96 concert | 0.96 length | 0.95 premiere | 0.95	DESIGN: watch | 0.95 leather | 0.95 pen | 0.95 pocket | 0.95 bag | 0.94	FASHION: adventure | 0.90 wallet | 0.88 tote | 0.87 enamel | 0.87 anti | 0.86
FILM & VIDEO: portrait | 0.93 documentary | 0.93 stretch | 0.92 funeral | 0.88 refugee | 0.88	FOOD: keto | 0.75 bitter | 0.72 knife | 0.71 butcher | 0.70 iconic | 0.69	GAMES: 28mm | 0.97 novel | 0.95 miniature | 0.94 5e | 0.94 visual | 0.93
JOURNALISM: winning | 0.77 annual | 0.76 och | 0.76 award | 0.75 edition | 0.72	MUSIC: folk | 0.94 heading | 0.94 alt | 0.93 printing | 0.93 roll | 0.92	PHOTOGRAPHY : muse | 0.88 audience | 0.88 mature | 0.88 monograph | 0.86 contain | 0.84
PUBLISHING : letterpress | 0.93 mountain | 0.92 ocean | 0.91 coast | 0.91 picture | 0.91	TECHNOLOGY : oscilloscope | 0.85 cortex | 0.82 toothbrush | 0.79 ruler | 0.79 nixie | 0.77	THEATER : identity | 0.88 installation | 0.87 produced | 0.87 satire | 0.87 cycle | 0.86

Rewards

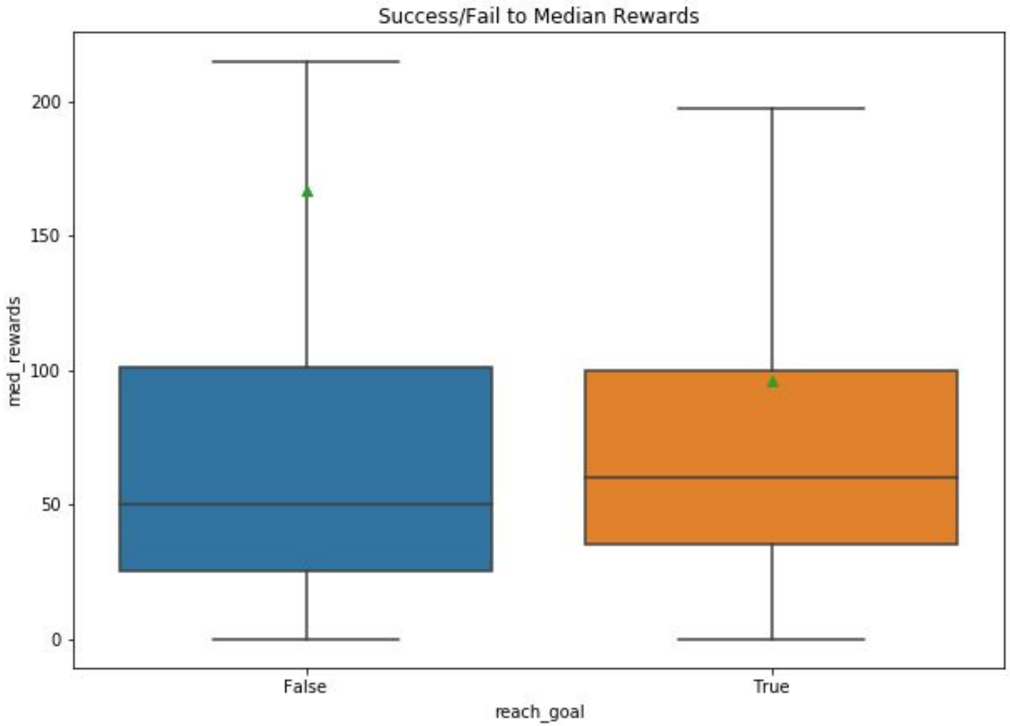
Looking at a zoomed in view of a scatter plot, we can see that past a certain combination of median rewards and goal do failed projects starter to appear.

At a median reward of $25, projects that have a goal of 10,000 tend to fail. At a median reward of $50, they tend of fail at higher goal levels.



Rewards

When looking at the boxplot, however, successful and failed projects tend to have the same range in their data. They are more affected by outliers since failed projects tend to have a higher, more unattainable, goal.



In-Depth Analysis of Machine Learning Models

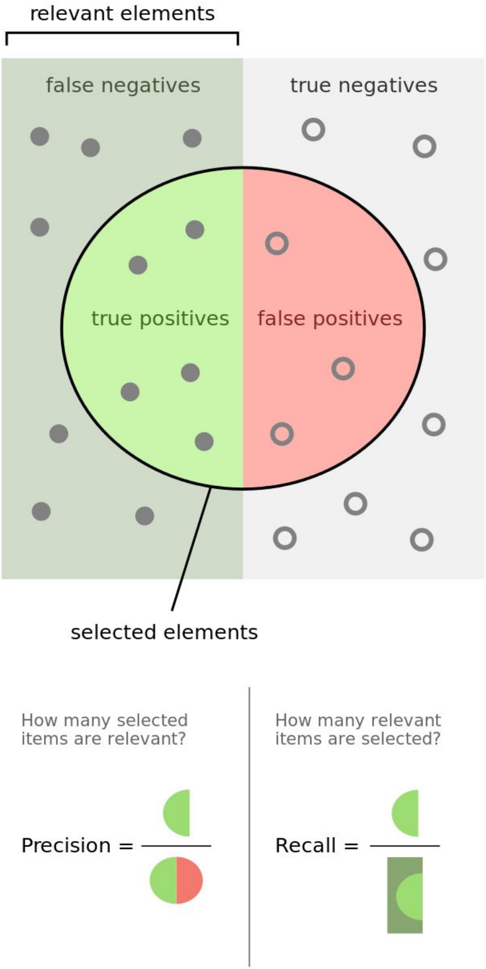
Scoring Method

Accuracy Score is a good base but does not address our business case

Focused mainly on predicting failure

Precision to improve confidence on determining failure

Use fbeta_score with a beta of 1.1 to move the weight towards recall



Base Models

We'll be exploring:

Logistic Regression

Linear SVC

K-Nearest Neighbors

Random Forests

Model	Score
Logistic Regression	69.63%
Linear SVC	68.92%
K-Nearest Neighbors	62.34%
Random Forests	68.56%

Tuning Hyperparameters

Focusing on Logistic Regression and Random Forests

GridSearchCV for Logistic Regression
- Score: 69.77%

RandomizedSearchCV for Random Forests
- Score: 72.75%

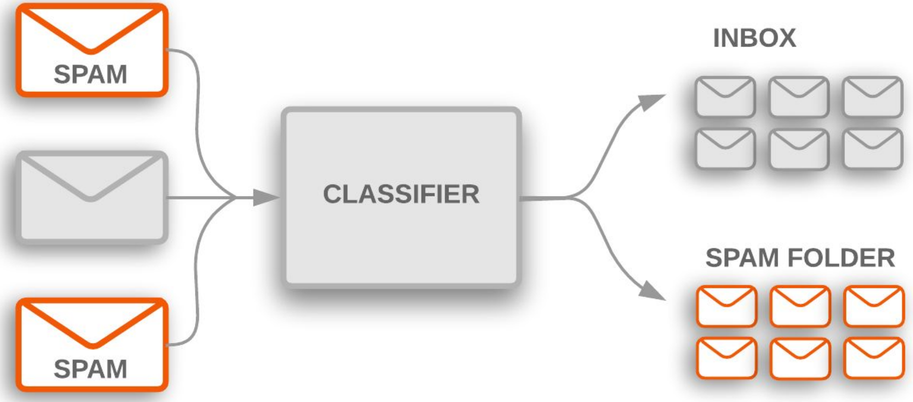
Logistic Regression Parameters Grid	
C	[0.0001, 0.001, 0.1, 1, 10, 100]
Random Forest Parameters Grid	
bootstrap	[True, False]
max_depth	[10, 12, 13, 16, 18, 21, 23, 25, 27, 30, None]
max_features	[‘auto’, ‘sqrt’]
min_samples_leaf	[1, 2, 4]
min_samples_split	[2, 5, 10]
n_estimators	[100, 200, 300, 400, 500]

Stacking Ensemble

Utilize MultinomialNB on blurbs and descriptions

Utilize multiple MultinomialNB classifiers for various scenarios of text analysis

Place predictions for each scenario as a new column in the dataset

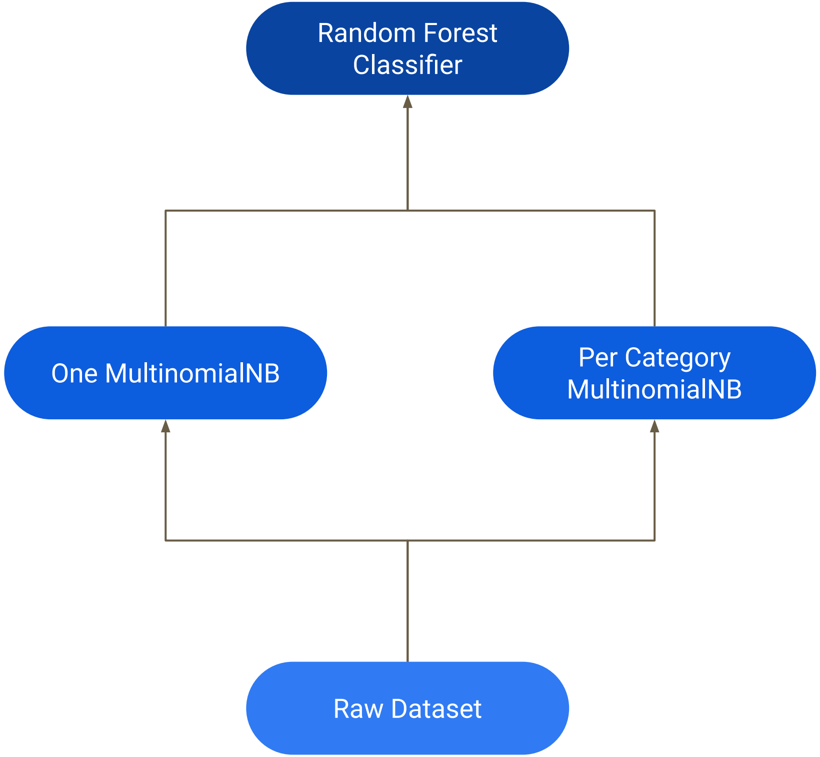


Stacking Ensemble

Test which set of columns performs the best

- One MultinomialNB: 74.50%

- Per Category MultinomialNB: 77.41%



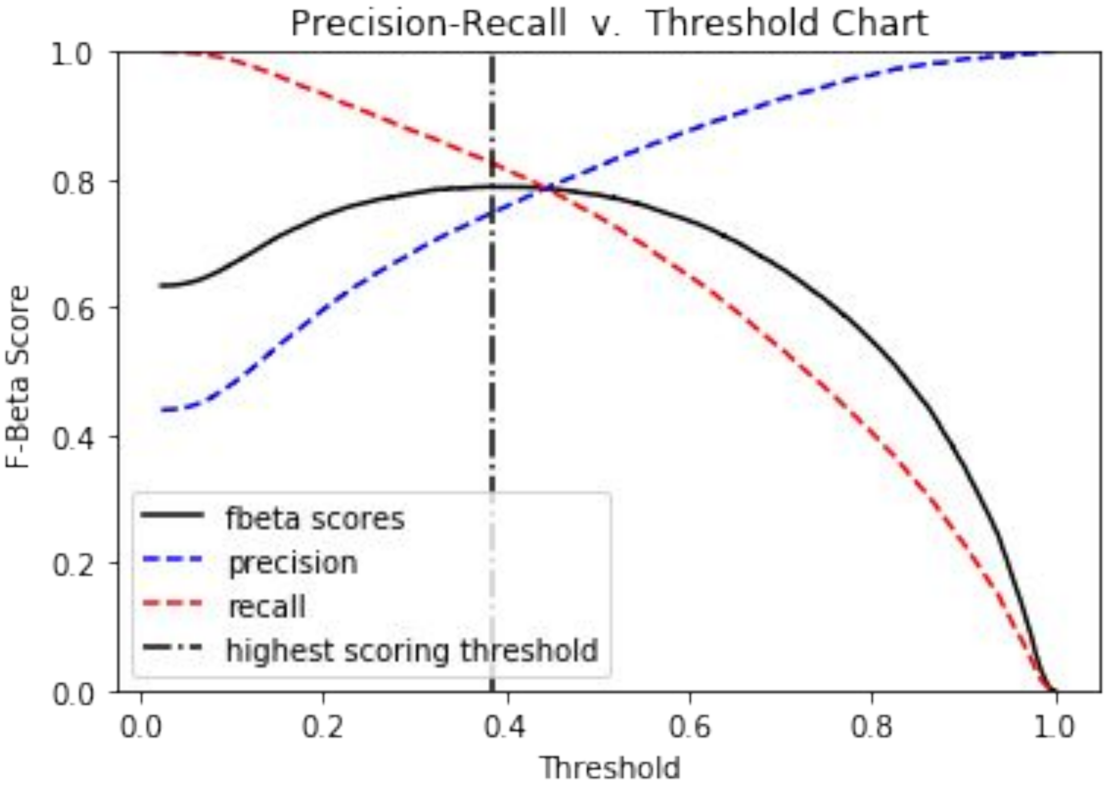
Thresholding

Determine which threshold, based on our beta of 1.1, gives the highest fbeta score.

Acquire precision, recall, and threshold through precision_recall_curve formula

Create a precision-recall v threshold graph

Best Threshold: 0.384
Highest Score: 78.84%



Final Model & Future Considerations

Our final model is a stacking ensemble:
- Per Category MultinomialNB classifier
- Tuned Random Forest classifier
- Adjusted Threshold to 0.384

A better defined text preprocessing system
- improved handling of foreign words
- improved lemmatization
- addressing proper nouns

A deeper analysis of dates

Recommendations

As creators, there are many factors to consider before employing a successful Kickstarter project.

The category can play a big factor as some are more likely to be successful.

Timing of projects can also play a big part. Projects finishing in January have the lowest success rate. Projects ending in April have the highest. Starting figures also affect project success.

Reward prices, along with project goal, can change how successful a Kickstarter is. Projects with a median reward of $25 and a goal of $10,000 tend to fail. Increasing the median reward amount and simultaneously decreasing the overall goal will increase chances of success.