Arianna Lang Wang
Springboard Jan 2nd, 2018 Cohort

Kickstarter Projects Written Report

April 30, 2018

DATA OVERVIEW

According to Wikipedia, “Kickstarter is an American public-benefit corporation that maintains a global crowdfunding platform focused on creativity. The company's stated mission is to ‘help bring creative projects to life’. People who back Kickstarter projects are offered tangible rewards or experiences in exchange for their pledges.This model traces its roots to subscription model of arts patronage, where artists would go directly to their audiences to fund their work.”

The dataset for this project is available on the Kaggle website: https://www.kaggle.com/kemical/kickstarter-projects/data. I only intend to use the year 2018 data set, which contains 378,661 rows and 15 columns. Each row is a record of a Kickstarter project and the columns contain a variety of information about that particular project such as the name of the project, launch date, goal amount, pledged amount, number of backers, the state of the project, etc.

You can access the python code I wrote for this project via my GitHub account: https://github.com/ariannalangwang/Capstone-Project-Kickstarter-Projects/blob/master/Kickstarter%20Projects%20Python%20Code.ipynb

CLIENT AND PROPOSED APPROACH

My client is someone who does not have much prior knowledge or experience in Kickstarter projects. He wants to gain some general understanding of the Kickstarter world and is also interested in potentially backing some projects himself.

One of the problems I'm trying to solve is to predict whether a Kickstarter project will succeed or fail. Another problem is determining how to make simple recommendations on which projects my client should back.

My proposed approach is to use graphs and charts to show an overall landscape of the Kickstarter world, use machine learning and natural language processing algorithms to predict whether a Kickstarter project will succeed or fail, then use what I learn to make recommendations on which projects are good to back.

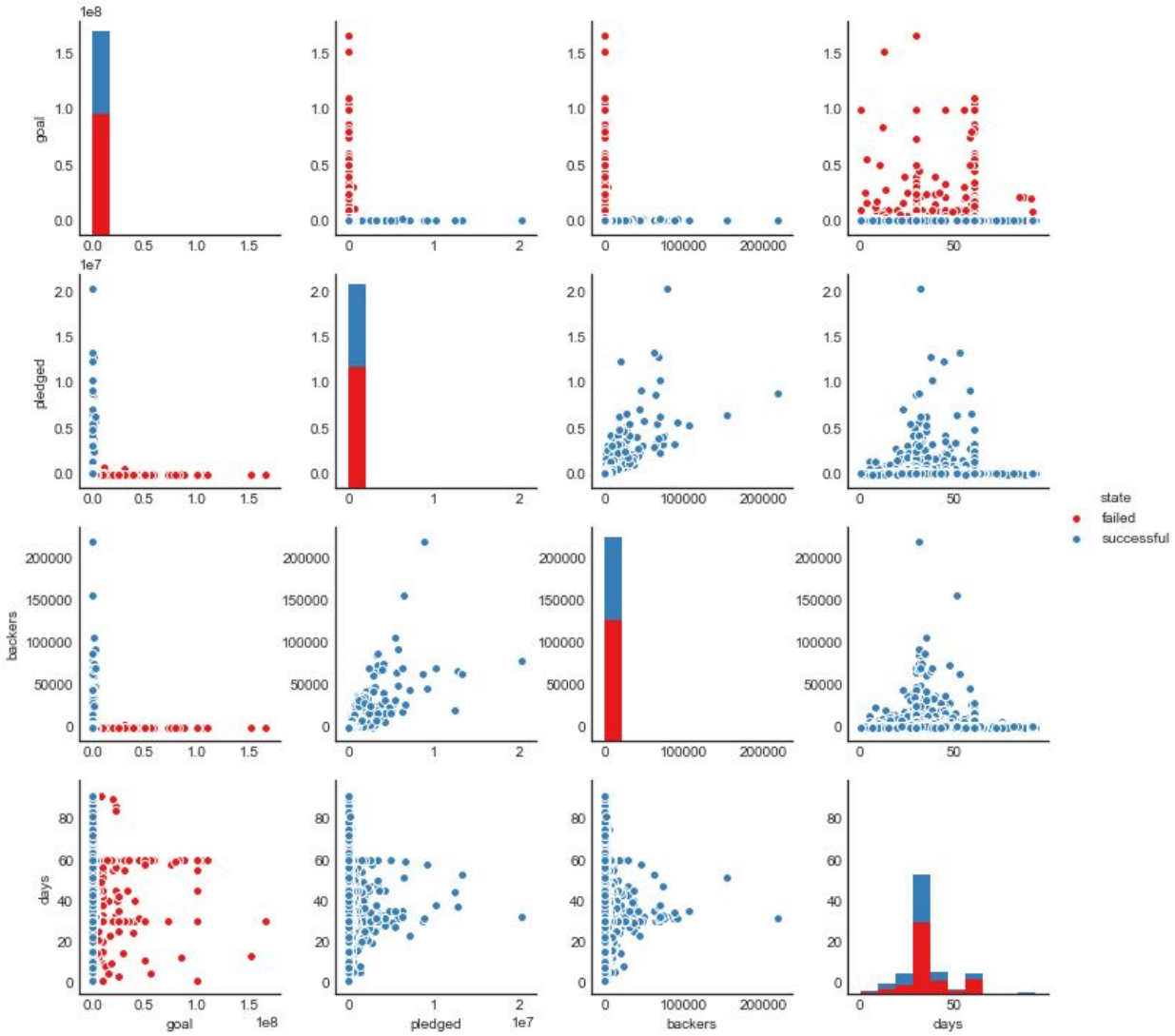
DATA CLEANING

The raw data was downloaded as a CSV file from the Kaggle website https://www.kaggle.com/kemical/kickstarter-projects/data. I used Pandas to import the CSV file as a data frame and named it df. I then immediately used df.drop_duplicates( ) to drop duplicate rows and grabbed the columns I wanted to keep in the data frame. I then used df.isnull( ).sum( ) to check missing values for each column. There were only four missing values in the 'name' column and I dropped them. I further changed a couple of column names and subset the data frame to include only rows where the 'state' column was either a 'success' or a 'failure'. There are 133,956 records of successful projects and 197,716 records of failed ones.

There are two columns named 'launched' and 'deadline' that contain date information. I first applied pd.to_datetime to change both columns into DateTime objects, then applied a function I wrote to extract date information from both columns. I then added a new column 'days' to the data frame, which is the elapsed number of days between launched date and deadline date.

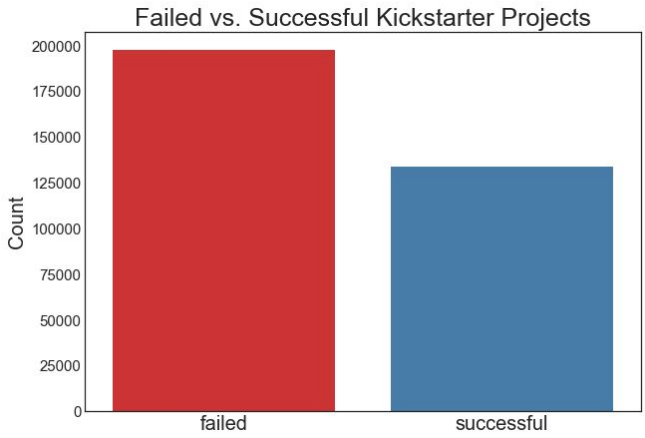
Another crucial step during data cleaning is to check for outliers. After using df.describe( ), I saw that four columns - 'goal', 'pledged', 'backers', and 'days' - have numeric values. I also realized that columns 'goal', 'pledged' and 'backers' have very large maximum values, which are indications of potential outliers. Drawing boxplots with df.plot.box( ) further proved my suspicion because the boxplots looked strange - individual points stacked on top of each other rather than forming boxes. I further investigated this with a pairplot. The pairplot proved that columns 'goal', 'pledged' and 'backers' have very skewed data.

One way to access outliers is to use the IQR (interquartile range) method. We first compute IQR = Q3 - Q1, then lower fence = Q1 - 1.5 * IQR and higher fence = Q3 + 1.5 * IQR. Any values outside the fences are considered outliers. After setting the outlier constant to 1,000 instead of the default 1.5, there were still a long list of values that were out of the fences. Clearly, using the IQR method to detect outliers is impractical for this case. The pairplot I did earlier clearly showed that columns 'goal', 'pledged' and 'backers' have very skewed data. This is the nature of this data set; hence, after some consideration, I decided to not eliminate any 'outliers' from this data set. The pairplot also made me realize that when I do exploratory analysis on some of the skewed-data columns, it is better to use the median instead of the mean as the central statistics measure.

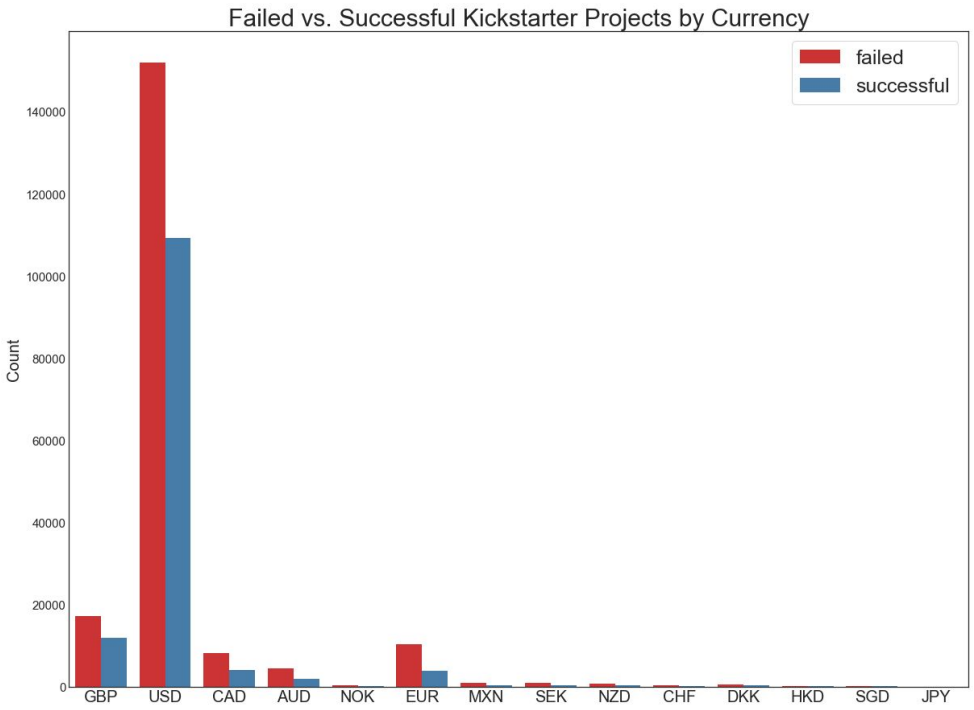


EXPLORATORY DATA ANALYSIS

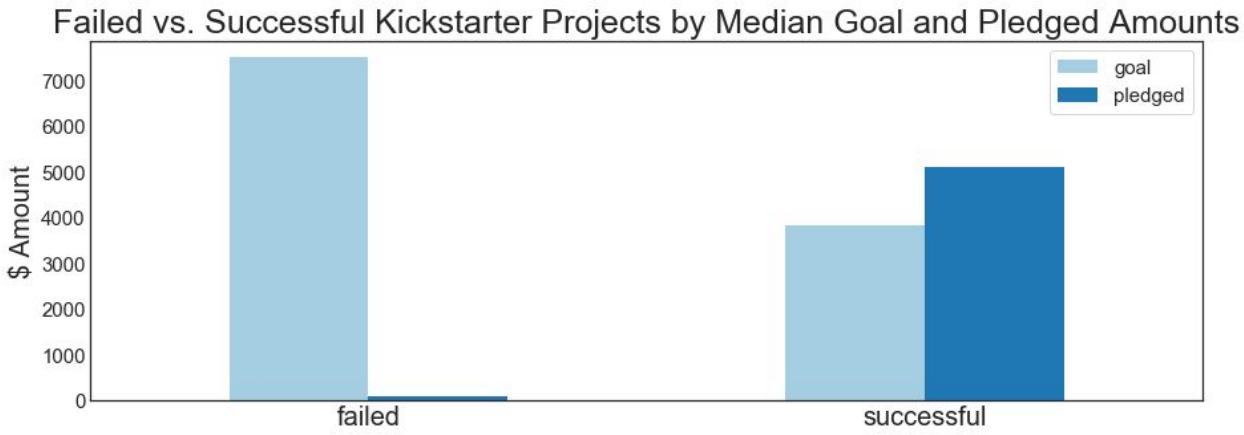
Since one of the important problems we want to solve is to predict whether a Kickstarter project will succeed or fail, it is a good idea to start some visualizations with 'failed projects' versus 'successful projects'. This first chart below shows the number of failed and successful projects, respectively. There are 197,716 failed projects versus 133,956 successful ones. Although there is a difference between the two numbers, the imbalance is not large enough to call for concern when I use machine learning for predictions.



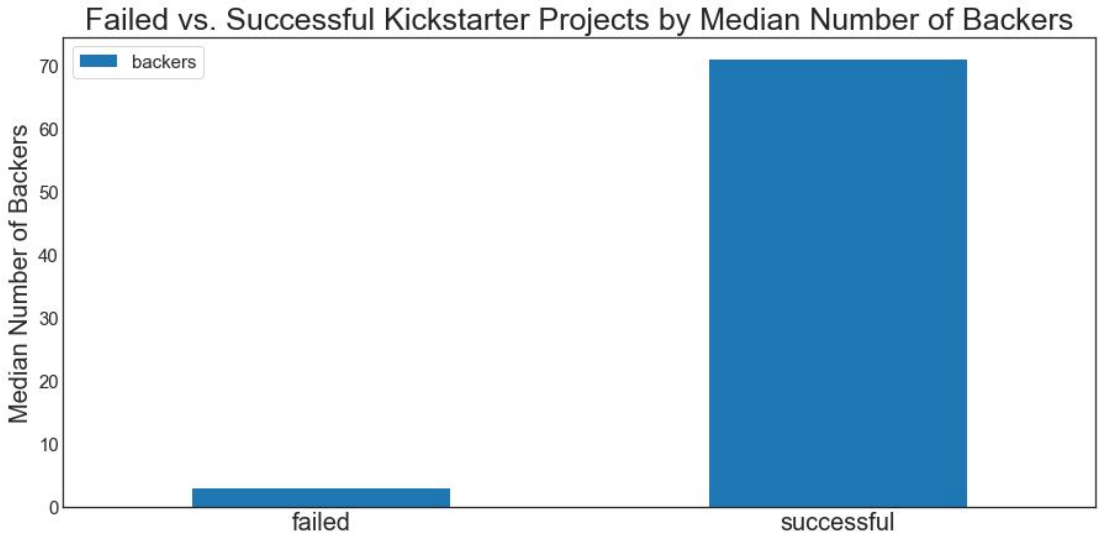
My client is based in the United States, he wants to know how many Kickstarter projects are based in U.S dollars. The second chart shows the number of failed and successful projects by currency. Most of the Kickstarter projects are in U.S dollars. This is not surprising since Kickstarter is an American company.



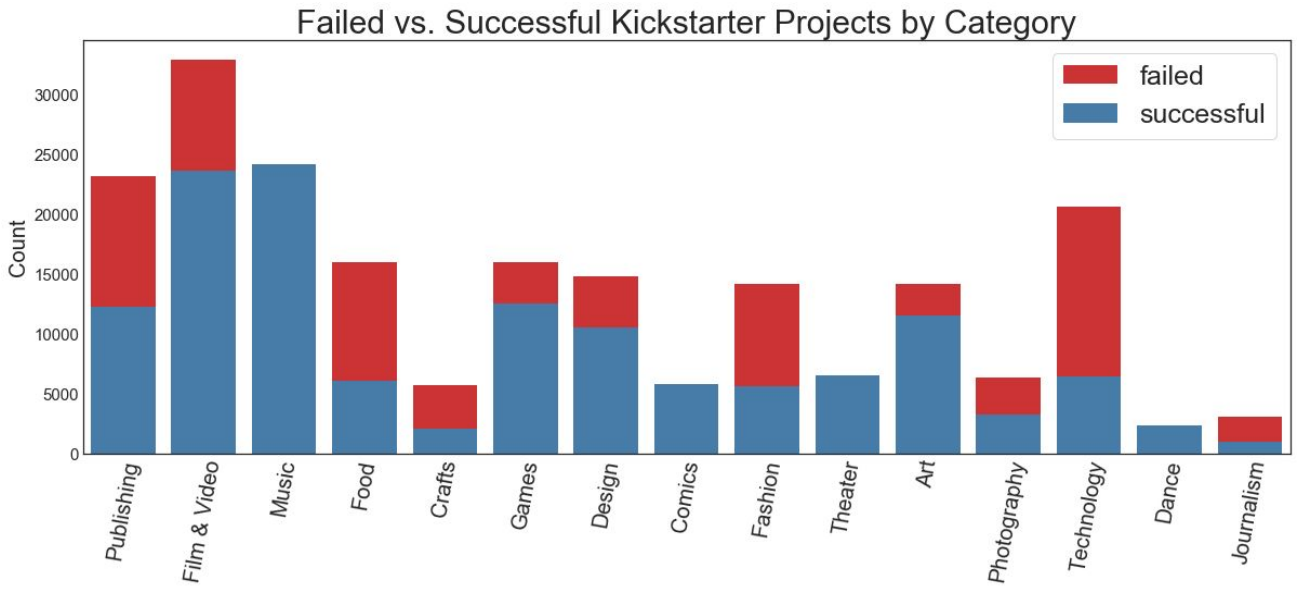
The third chart shows the median goal and pledged amounts in U.S. dollar equivalents for failed and successful projects. It's interesting to note that the median goal amount of failed projects is much higher than that of successful ones.



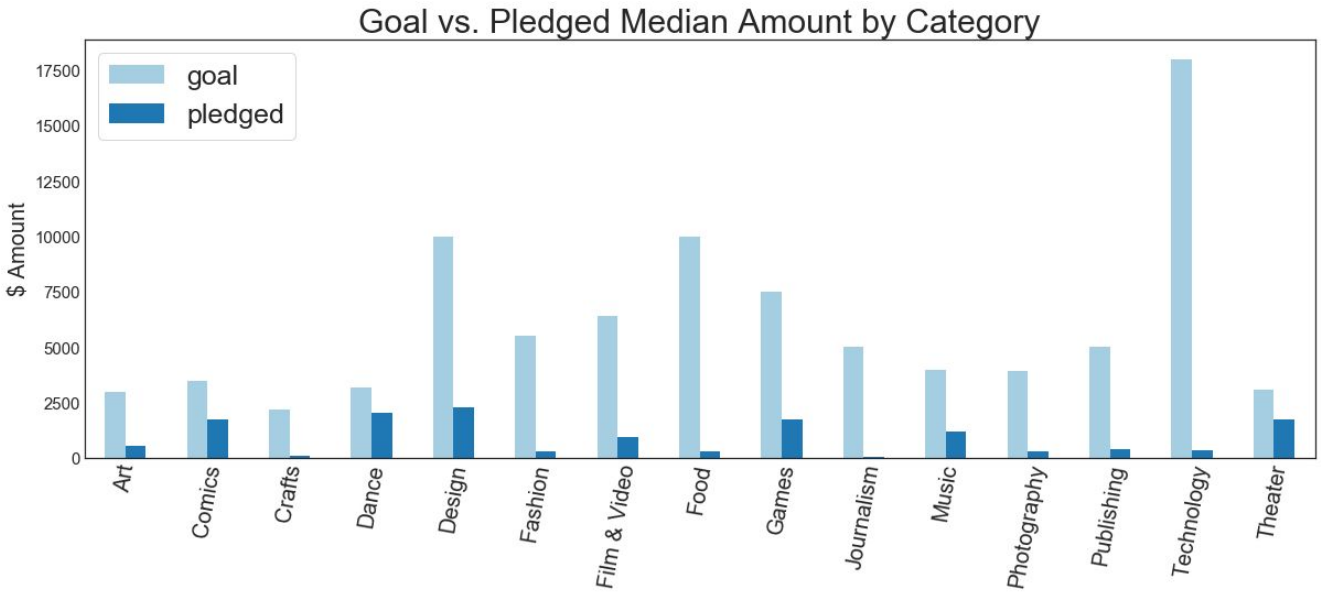
The fourth chart shows the median number of backers of failed and successful projects. As expected, the median number of backers of successful projects is much higher than that of failed ones.



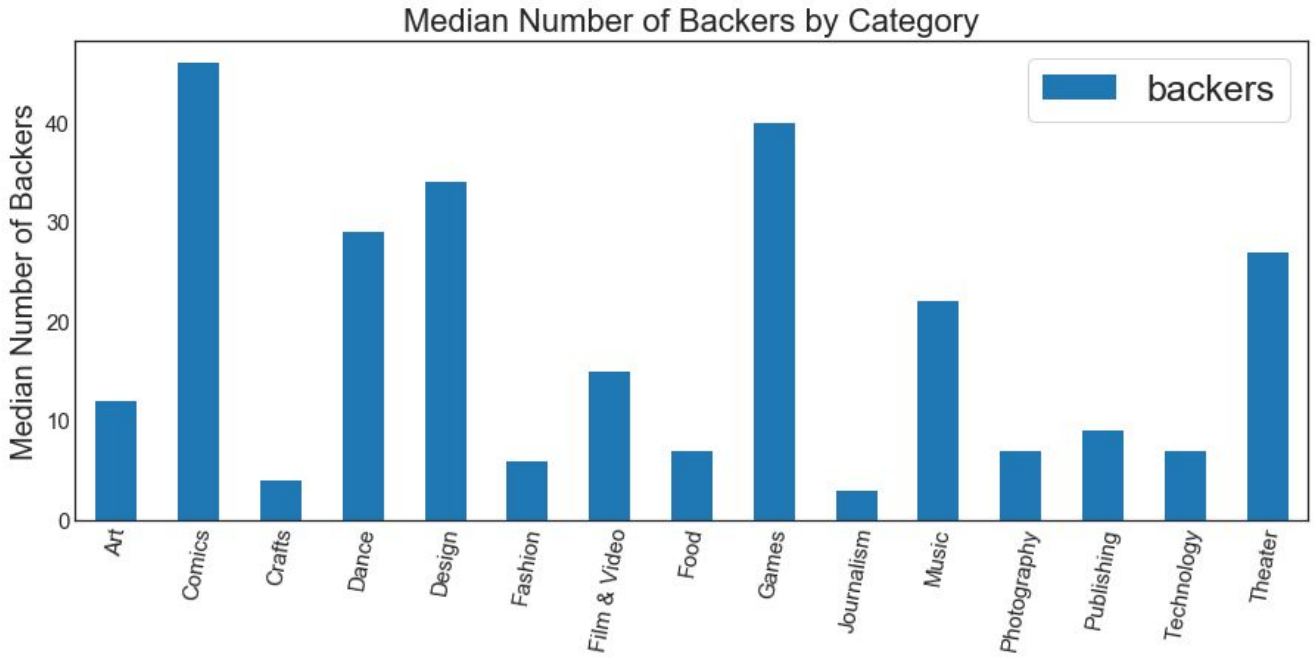
My client is interested in backing projects in certain categories; therefore, it is also a good idea to show some visualizations by category. The fifth chart below shows the number of failed and successful projects by category. We see that the film & video category has the most number of Kickstarter projects, followed by the music category. We can also see that so far in 2018, none of the projects in the music category have failed. Projects in the art and game categories also have high probabilities of success.



The sixth chart below shows the median goal and pledged amounts in U.S. dollar equivalents for projects by category. The median goal amounts are high for the technology category, but the actual median pledged amount for the same category is low.



The seventh chart below shows the median number of backers for projects in each category. The comics and games categories have the highest median number of backers.



PREDICTIONS USING MACHINE LEARNING

After that basic understanding of the overall landscape of the Kickstarter project world, we can use machine learning algorithms to predict whether a Kickstarter project will succeed or fail. One of the columns in the data set is called 'state'. This column has binary values of 1 and 0, with 1 being a 'successful' Kickstarter project and 0 being a 'failed' one. The 'state' column is the target. The features to feed into the machine learning algorithms are the goal amount in U.S. dollar equivalents, the number of backers, the number of days elapsed between launch date and the deadline, categories and currency. A quick heatmap shows that there will be no problems with multicollinearity because the correlations between each pair of the features are low.

Before fitting the data, a good question to ask is, "Are all the features important for making predictions?" We can use a random forest model to evaluate feature importance on a classification task. The result shows that the goal amount in U.S. dollar equivalents, the number of backers and the number of days elapsed between launch date and the deadline are the most important features for making a prediction. We will

use only these three most important features to fit our models. The precision and recall rates will be lower than fitting the model with more features, but we save run time on computation.

I fitted the data with two different models: random forest and logistic regression. It is very computationally-expensive to do a randomized or full grid search with my laptop, so I will fit the data with default model parameters only. Please note that I demonstrated how to do a grid search for optimal parameters in my previous capstone project: Global Terrorism Analysis. It is also available on my GitHub account: https://github.com/ariannalangwang/Capstone-Project-Global-Terrorism.

The classification table for the random forest model is as follows:

	precision	recall	f1-score	support
0	0.93	0.93	0.93	197716
1	0.89	0.89	0.89	133956
avg / total	0.91	0.91	0.91	331672

Here is the classification table for the logistic regression model:

	precision	recall	f1-score	support
0	0.89	0.96	0.92	59400
1	0.93	0.83	0.88	40102
avg / total	0.91	0.91	0.91	99502

We're interested in the precision rate and the recall rate of 'successful' projects. They are represented by the row that is labeled as "1" in the classification tables.

In statistical terms:

precision = true positive / (true positive + false positive) = true positive / positive predictions

recall = true positive / (true positive + false negative) = true positive / positive true state

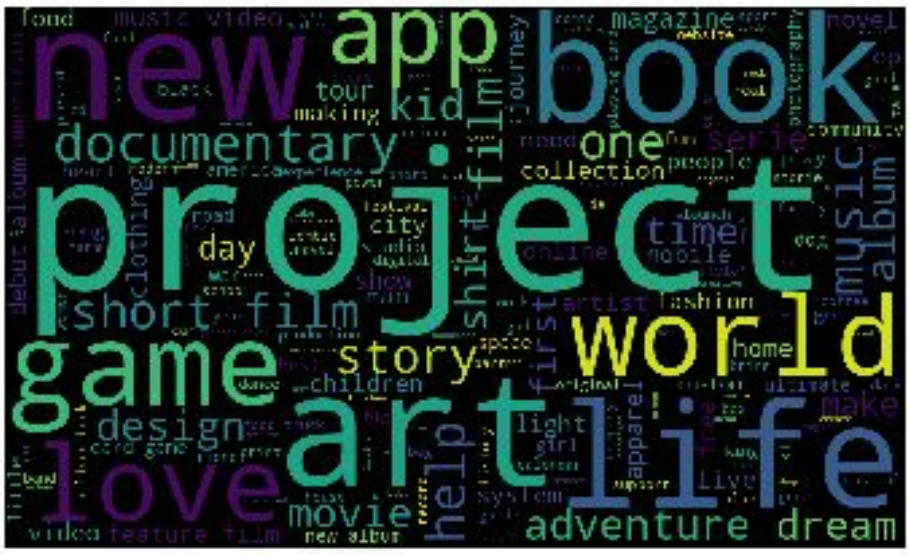
In our case, precision is the percentage of the predicted successful projects that are actually successful. Recall is the percentage of the successful projects that are predicted correctly by our model. To measure a models' performance, we want both rates to be as close to 1 as possible with 1 being the highest.

From the above classification tables, we can see that the logistic regression model has a higher precision rate (0.93) for 'successful' Kickstarter projects and the random forest model has a higher recall rate (0.89) for 'successful' projects. Depending on which rate we're interested in using, we can choose either of these two models.

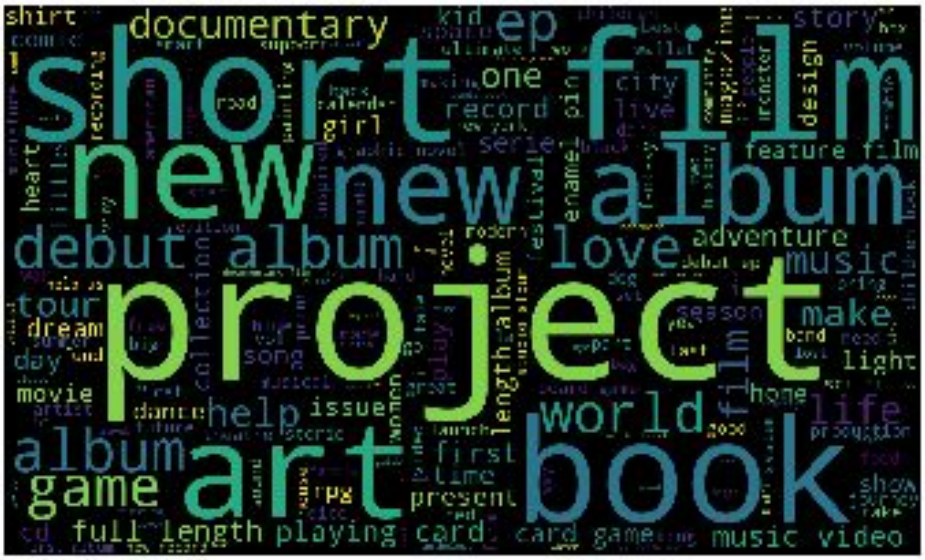
TEXT CLASSIFICATION USING NLP

The original dataset has a text-based column called 'name'. Can I predict whether a Kickstarter project will succeed or fail solely based on this 'name' column? To visualize some of the frequent words used as the names of the projects, I made two word clouds;

one for all of the failed projects:



and one for all of the successful projects:



I notice that the word ‘album’ appears at least three times in the successful projects’ word cloud, but not even once in the failed projects’ word cloud. This result coincides with a previous chart I did on the number of failed versus successful projects by category. In that chart, none of the projects failed in the music category.

Before using classification models for the prediction, I used a Tf-idf Vectorizer to convert text-based data into a weighted numeric matrix. Tf–idf, short for 'term frequency–inverse document frequency', is a numerical statistic that is intended to reflect how important a word is to a document in a corpus. The tf-idf value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some common English words such as 'the', 'a', appear more frequently in general. After having the text-based feature converted into a weighted numeric matrix, I then use either a Naive Bayes classification or a Random Forest classification to do the predictions.

The classification table for the Naive Bayes model is as follows:

	precision	recall	f1-score	support
0	0.65	0.88	0.75	197716
1	0.63	0.29	0.40	133956
avg / total	0.64	0.64	0.61	331672

The classification table for the random forest model is as follows:

	precision	recall	f1-score	support
0	0.66	0.81	0.72	59400
1	0.56	0.37	0.45	40102
avg / total	0.62	0.63	0.61	99502

As expected, predicting whether a Kickstarter project will succeed or fail solely based on the name of the project does not have as good a predictive power as compared to what we've done earlier with other features.

RECOMMENDATIONS FOR BACKING A NEW PROJECT

My client is interested in backing some Kickstarter projects in the film & video category. He wants some recommendations on which project he should back.

Since the original data set has the goal amount and the pledged amount in U.S. dollar equivalents for each project, we can compute a new variable 'surplus', which is the pledged amount minus the goal amount. Any project with the surplus greater than zero will be a successful project to back. After subsetting the data set with only 'live' projects, i.e. projects that are open and accepting backings, I sorted the 'live' projects from the highest to lowest surplus amounts and printed out the top twenty projects. Here are the results:

	name	category	main_category	goal	pledged	backers	state	surplus
319442	MADE YOU LOOK	Shorts	Film & Video	3500.00	6204.00	81	live	2704.00
242907	The Whibbits! A stop motion animated web serie...	Animation	Film & Video	818.39	2482.44	30	live	1664.05
190837	LaVoy: Dead Man Talking	Documentary	Film & Video	12000.00	13450.00	125	live	1450.00
341218	CADABRA - A Short Film	Comedy	Film & Video	406.55	1599.09	7	live	1192.54
127730	The Walk	Film & Video	Film & Video	6775.80	7479.13	73	live	703.33
132682	"BRIDGES" a short doc about the in-betweeners	Documentary	Film & Video	7000.00	7656.00	56	live	656.00
314947	Lady (short film)	Shorts	Film & Video	1295.78	1853.65	36	live	557.87
127532	Code Switch	Horror	Film & Video	3000.00	3511.00	27	live	511.00
246730	FATHOM | an Eco-Thriller	Shorts	Film & Video	6000.00	6452.00	35	live	452.00
43336	BEAT IT	Shorts	Film & Video	1355.16	1775.26	20	live	420.10
172567	Demon - A Short Film	Horror	Film & Video	100.00	430.00	6	live	330.00
14490	Animated TV Pilot	Animation	Film & Video	7500.00	7817.00	49	live	317.00
86009	Nightwatcher	Shorts	Film & Video	3150.00	3450.00	80	live	300.00
111007	The Yellow Wallpaper	Shorts	Film & Video	2000.00	2280.00	28	live	280.00
113526	I'm going to Paris...	Drama	Film & Video	7000.00	7270.00	39	live	270.00
261103	Turntable (A Short Film)	Drama	Film & Video	2000.00	2240.00	19	live	240.00
267746	Angela Plays Ukulele Original Music Video Funding	Music Videos	Film & Video	100.00	212.00	7	live	112.00
222587	Al Hmar - A Short Phoenician Dark Comedy [Post...	Comedy	Film & Video	6000.00	6068.00	37	live	68.00
224218	Black History Month - A Blaxploitation Film	Action	Film & Video	100.00	160.00	3	live	60.00
368967	Seeking Sole Mates	Shorts	Film & Video	15000.00	15045.00	56	live	45.00

These twenty live projects are my recommended projects to back because no matter which one(s) my client chooses, it is guaranteed that these projects will be successful.

FINAL WORDS

This project has a lot of room for further analyses and fine-tuning. My exploratory analyses could be more in-depth instead of merely stating obvious observations for each graph. Because of the computational-power constraint, I did not run a random or full grid search to find the best parameters for my machine learning models. Furthermore, I could have used the random forest and logistic models I trained earlier to make predictions on which 'live' projects will be successful and then make recommendations to my client this way. Because of the time constraint on this project, I will stop here, for now.