Capstone Project 1 Kickstarter Projects Success Prediction Milestone report

Zongkai Wu

1. The problem:

Kickstarter is a crowd-fund based website for creators to raise funds for their projects and in return provide their backers exclusive rewards and benefits. In this project, we are going to analyze previous kickstarter projects and provide predictions whether a new project will be successful or not.

The potential clients of this project include: kickstarter project creators, kickstarter backers and kickstarter company or similar crowd-funding companies.

For kickstarter project creators, by applying the most successful strategy from previous project creators, they will be more likely to meet the goal and raise as much funding as they needed to successfully finish their project.

For kickstarter backers, they could predict and analyze if a project they wanted to back are more likely to be successful or not, in return guide their decision of whether to back this project or not, and how much they should pledge.

For kickstarter company or similar crowd-funding companies, their revenue is directly decided by the success rate of their projects, since they take a 5% fee for every successful project. So for these companies, having a higher success rate will benefit both their revenue and attract more content creators to use their platform. Therefore, they can develop the most successful strategy to help increase the success rate on their platform. Focus their limited promotion resources on the projects that are most successful or need the help most to meet the goal.

2. Dataset:

The dataset we used in this project is from Web Robots, which uses a scraper robot to crawl all Kickstarter projects and collects data in CSV and Json formats. From March 2016 they run this data crawl once a month. In the project, we used the dataset from scrape date 2019-10-17, which is the most current dataset at the time this project started. Links to Web Robots kickstarter datasets: https://webrobots.io/kickstarter-datasets/

3. Initial findings:

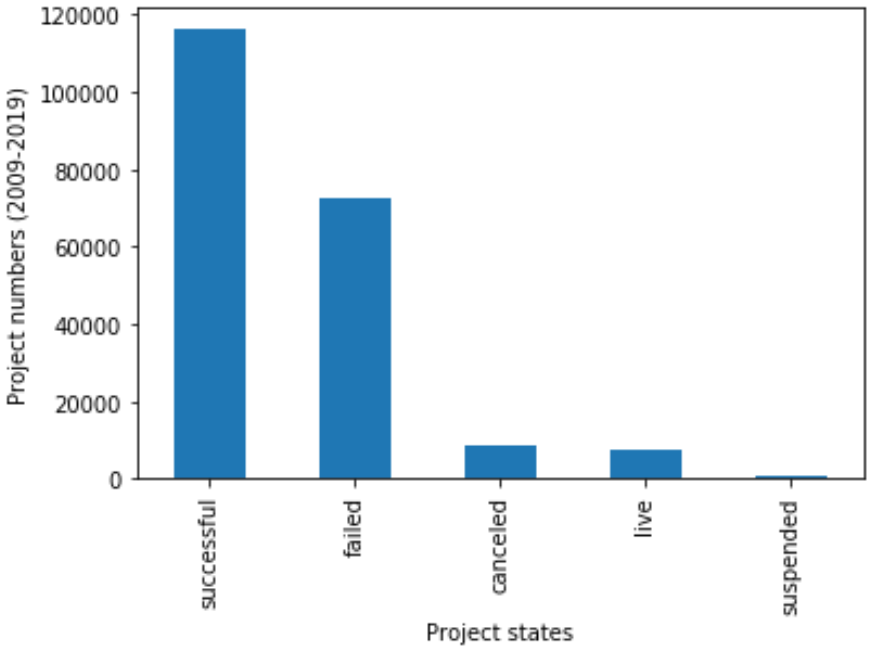

Figure 1. Project states distribution histogram

From the state distribution histogram, we can see there are more projects successful compared to failure, and canceled, suspended and live projects were much smaller compared to finished projects considering we included projects for five years, and most projects only last for 1 month or 2 month. Also there are 11k successful projects and 7k failure projects, which provide enough entries for our analysis and train models.

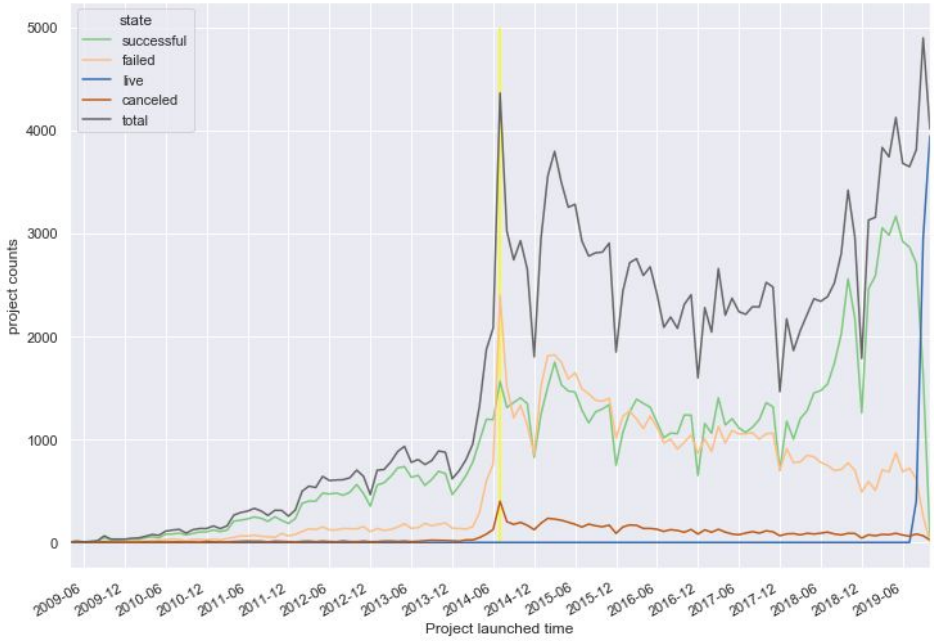

Figure 2. Time series graph with project counts with different states

From Figure 2, we notice a few interesting patterns for the project launch time series. 1) There is a clear project count increase around July of 2014 (labeled by yellow line). We couldn't find a record of what happened to the kickstarter website that month, but our suspicion is there was a large promotion that largely increased their traffic. It can also be seen that the successful project number and failed project number both showed a peak at this month, but it's more obvious for the failed case than successful. This indicates although the promotion at that month boost website traffic, it sacrifices success rate a bit at the same time. 2) There's a clear cycle pattern with project counts, each December launched project will show a clear drop, while at the start of year project numbers increase significantly, and peak around June or July. This pattern is clearer to total and successful project counts, and can still be seen in failed counts.

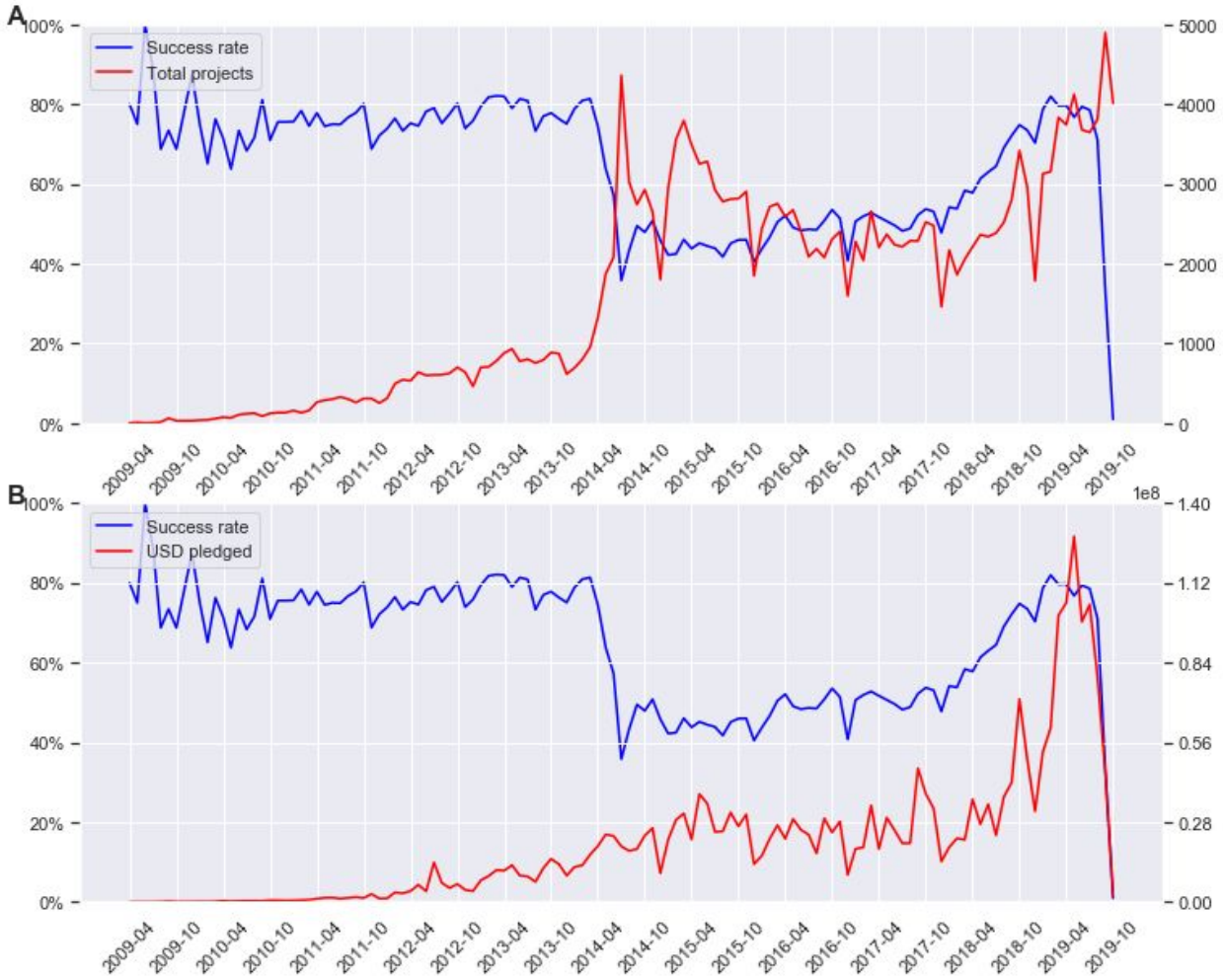

Figure 3. Time series A) Success rate compared to total project counts B) success rate compared to USD pledged

From Figure 3A, we got a similar conclusion from Figure 2, that during 2014-July, there's a clear increase in total project number, but decrease in project success rate, and success rate

gradually increase and get closer to previous level (~80%) in 2019. From Figure 3B, we can get an important message that although the success rate decreased a lot by 2014-July, the actual total USD pledged is actually increased due to the total project number increase. This value is more important for company revenue compared to success rate. Total USD pledged numbers remain relatively stable until 2018-Nov, which starts to see large increases again.

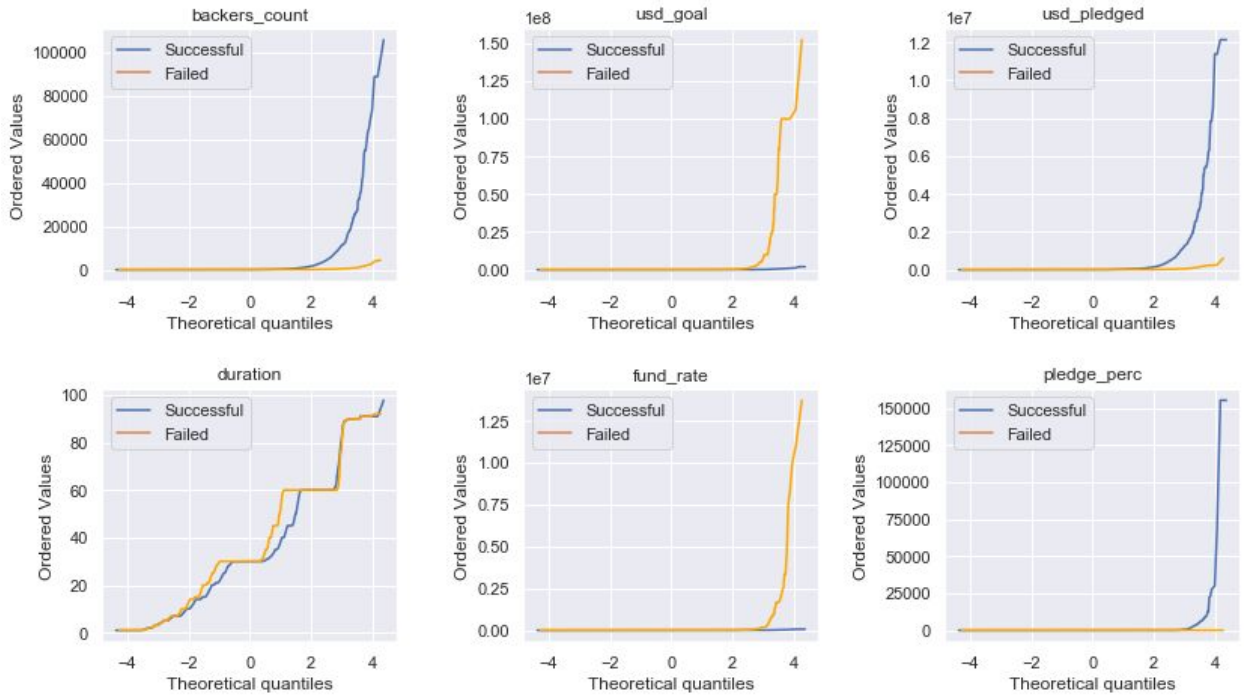

Figure 4. Probability plots for numeric columns

Quick histogram check for all numeric column distribution conditions. The result is not very clear because some columns have some far off outliers.I think use probability plots could show distribution clearer. Probability plot shows why for most numeric columns, histogram is not working.For most of the columns data, there's a very large range but most of the data (within 2 standard deviation, around 97%) are much small compared to the large data.Some interesting observation here is for fund rate and fund goal, failed projects tend to require higher amount and fund rate, which indicate it's more difficult for them to reach goal within target time. As for duration, there's no clear difference between successful and failed cases.

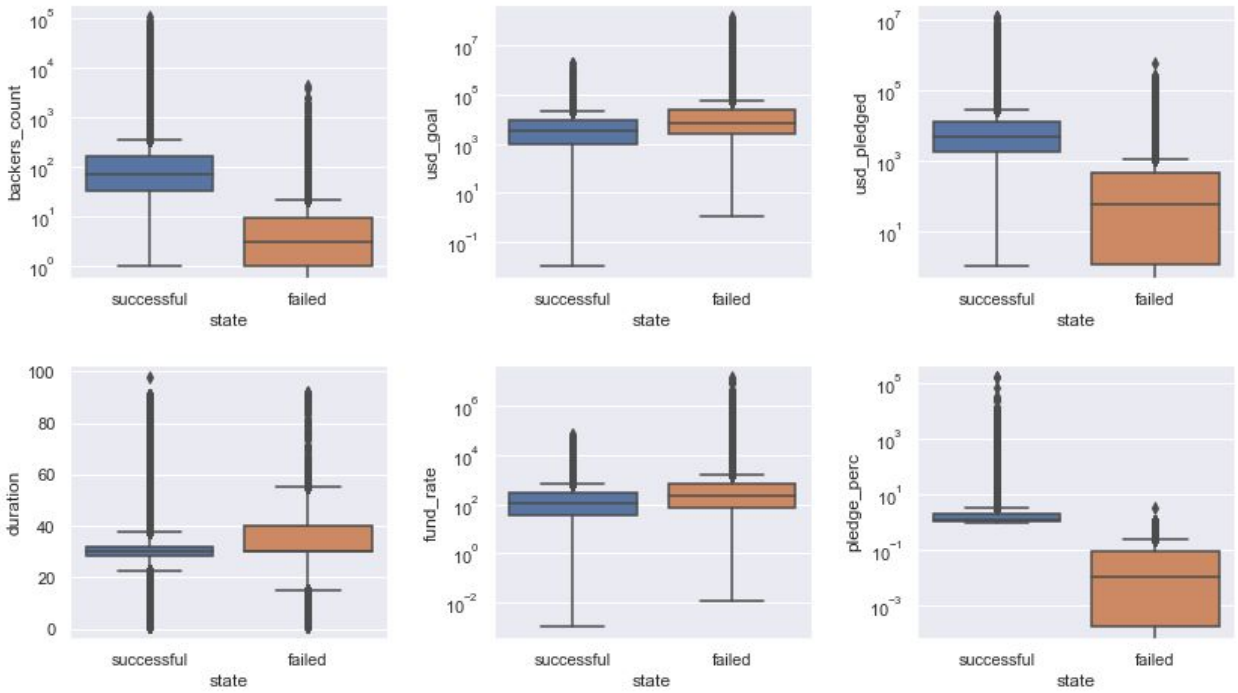

Figure 5. Box plots for numeric columns

Probability plot successfully shows the position of outliers and how far they can go for different states, but it doesn't provide a clear comparison due to the majority of the data being too small compared to the large outlier. So we use box plots, and take log scales for some data to show clear comparison. From these box plots, we can get several interesting conclusions, such as successful projects have more backers compared to failed one. Failed cases are likely to set higher goals, and interestingly also like to require longer time for funding. While projects with less fund request and shorter request time are more likely to be successful.

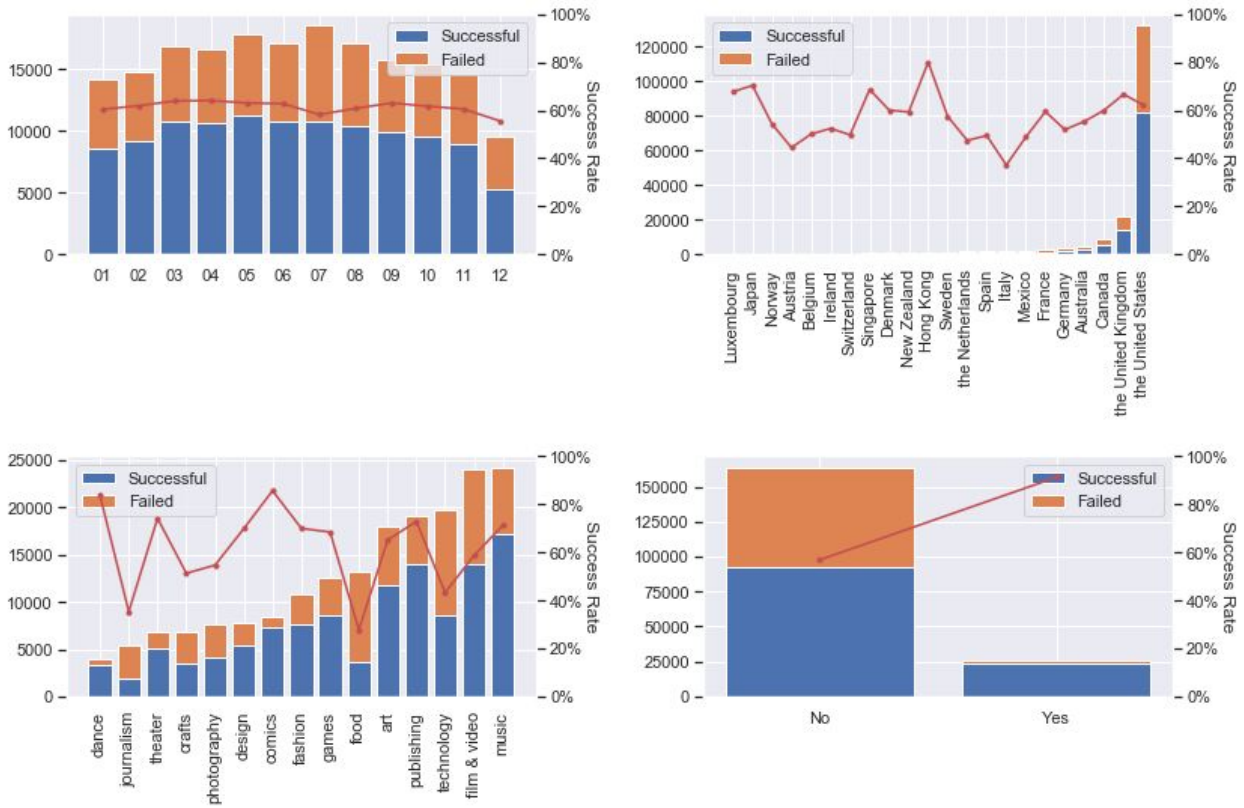

Figure 6. Stack bar plot for category columns

We use stacked bar plots to analyze category columns. From this figure we can see December has a smaller amount of projects launched compared to other months, while July has the most. Interestingly, the success rate was pretty stable but shows lowest at July and December, which are the months with most and least projects. For country distribution, most of the projects are from the US, second place is the UK, which only has around 1/6 of US project numbers. Third place is Canada, all other countries' project numbers are much smaller compared to the top 3.