Kickstarter Campaigns

Success Factors and Analysis

Monica Boomgaarden
College of Engineering and Applied Science
University of Colorado Boulder
Boulder CO United States
Monica.Boomgaarden@colorado.edu

Rodrigo Garcia
College of Engineering and Applied Science
University of Colorado Boulder
Boulder CO United States
Rodrigo.Garcia-1@colorado.edu

Nicole Lincoln
College of Engineering and Applied Science
University of Colorado Boulder
Boulder CO United States
Nicole.Lincoln@colorado.edu

1 INTRODUCTION

1.1 Problem Statement

Kickstarter is a website which first launched on April 28th, 2009 with the main purpose of bringing user inspired projects to life through community backing. It allows community members to provide financial support for a project's development and production. The final result of the project is then shared with the people who backed it. Since its inception, millions of people have provided support for the new ideas on Kickstarter's platform; resulting in approximately $5.3 Billion dollars being raised. The projects span a wide range of categories from Arts and Games to Dancing and Technology. Kickstarter's approach to project funding is "all or nothing". Therefore, if a project fails to achieve its community-based funding goals it receives no funding at all. If a project campaign fails, each backer's financial contribution is returned. As a result, achieving the desired funding goals for the project is crucial. Therefore with our analysis we explore the questions: Can attributes be identified which impact the outcome of a Kickstarter campaign? If so, which factors contribute most to the success of a Kickstarter project's success or failure? Addressing these questions could provide useful data for future projects success in achieving funding.

1.2 CCS Concepts

• Computing Methodologies • Machine Learning
• Data Science

1.3 Keywords

Kaggle, Machine Learning, Kickstarter, Success Prediction, Crowd Sourced Funding, Data Science, Logistic Regression

2 LITERATURE SURVEY

Kickstarter's results have been previously analyzed in other work by data scientists. We identified three articles that we thought would provide useful insights for our project. Each article discusses its own approach to analyzing the attributes present in its respective Kickstarter data, with the objective of developing a prediction model to determine the factors which contribute to a campaign's success. These articles were as follows:

- **[1] "We Analyzed 331,000 Kickstarter Projects. Here's What we Learned About Kickstarter Success"**: This article's primary goal was to look at the relationships between categories, subcategories and days of the week and having a successful funding campaign.
- **[2]** "**Kickstarter Analysis and Prediction"**: This paper looked at the category, goal, and country and attempted to create a predictive model of whether or not a project would meet its funding goals.
- **[3]** "**Using Machine Learning To Predict Kickstarter Success"**: This project attempted to create a predictive model for a Kickstarter project's success or failure by analyzing a

wide range of attributes. They found the following attributes to have the most impact: goal, length of campaign, month campaign was started in, time of launch, country of origin, category, day of the week, and being a staff pick.

Overall, none of the articles we reviewed found any single attribute to be a strong predictor of a Kickstarter project's success. Additionally, the predictive models created seemed to give between 60-70% accuracy at best. Logistic regression seemed to be the model which provided the most success.

3 PROPOSED WORK

3.1 Preprocessing

It appears that the majority of the data is in order requiring little if any data clearing. With that said, a detailed data quality assessment will still be performed by leveraging python to detect and identify rouge or empty data values with some basic exploratory data analysis. Any missing or malformed data values will be replaced with each attribute's mean value. We will also need to examine the data for statistical outliers that may impact our results. Additionally, we can reduce some attributes by evaluating and using the most accurate values for US Dollar Conversion.

We will also need to perform some discretization and transformation of some of the numerical values such as the monetary values, including goal and pledged amounts, the date for launch and the number of backers. In order to use some of the methods that are available in the Python libraries, we will also need to normalize the data. We will likely use the z-score normalization for this.

3.2 Process For Derived Data

Once we have performed the preprocessing of the data, we will perform some exploratory data analysis, using several of the available libraries in Python, including Seaborn, which offers a variety of statistical visualizations.

We will be looking to see if we find similar results to those of the analysis by Laura Lewis. Lewis's analysis identified several positive and negative influences on the likelihood of a campaign reaching its goal. However, Lewis's dataset only included just over 209K projects. Our dataset is a bit larger, as we anticipate having over 330K projects to analyze after data cleaning.

As discussed below, we anticipate preparing models using logistic regression and/or aive Bayes as classifiers, in order to predict whether a campaign is successful. We will be using additional visualizations, such as correlation heat maps, confusion matrices, and ROC curves, in our analysis in order to be able to visually analyze the results from our model.

4 DATASET

Our data set can be found at https://www.kaggle.com/kemical/Kickstarter-projects. There are two versions of the Kickstarter data available. We will be using the 2018 version, since it includes more complete and the dataset is more up to date. This is a multivariate dataset that provides information regarding the success and failure of recent Kickstarter projects. There are 15 attributes and 378,662 items in the dataset. The attributes include the following:

- **ID**: Numerical, unique campaign identifier
- **Name**: nominal, describes the name of the projected
- **Category**: categorical, actually a subcategory of the Main_category
- **Main_category**: categorical, describes the overarching category of the campaign (e.g. Music, Crafts, Games, etc.)
- **Currency:** categorical, describes the currency types for the monetary columns
- **Deadline:** Deadline for pledges to the campaign.
- **Goal:** numerical, ratio-scaled; monetary value that needed to be reached in order for campaign to be funded

- **Launched**: numerical, interval-scaled; date the campaign was launched
- **Pledged**: numerical, ratio-scaled; monetary amount that backers of the campaign agreed to pay
- **State**: Categorical, describes the current state of the project - successful, failed, or cancelled
- **Backers:** numerical; indicates the number of people who pledged support to the campaign
- **Country:** categorical; country location of the company that is seeking funding
- **usd pledged**: numerical; dollar amount of pledges converted into US dollars by Kickstarter
- **Usd_pledged_real:** numerical; dollar amount of pledges converted into US dollars by fixer.io
- **Usd_goal_real:** numerical; dollar amount of goal converted into US dollars by fixer.io

5 EVALUATION

We will be using logistic regression or naive Bayes to evaluate our models. The paper by Laura Lewis, cites extremely marginal if not slightly worse modeling outcomes using some other advanced prediction methods such as Principal Component Analysis, Random Forest, and XGBoost. These aforementioned methods also incur a substantial computational penalty; taking between 1.5x and 19x longer than logistic regression. Thus we will be using logistic regression or naive Bayes or both. We may assess other models if time permits.

6 TOOLS

To achieve our analysis and modeling goals, we have selected the following tools:

- Kaggle Notebooks
- Python and its relevant libraries.
- Tableau for presentation purposes, if necessary
- Powerpoint for project presentation
- Video presentation software will use a combination of iMovie for post-production, and quicktime or some other form of desktop video recording program.

7 MILESTONES

7.1 Todo

The milestones for our project continue to be as follows:

- December 4, 2020: Modeling
- December 7, 2020: Final Analysis and Report
- December 9, 2020: Project Presentation
- December 11, 2020: Peer Evaluation and Interview Questions

7.2 Completed

The following milestones for our project have been completed.

- November 13, 2020: Preprocessing
- November 20, 2020: Exploratory Data Analysis

8 CURRENT RESULTS

8.1 Preprocessing

Preprocessing for the kickstarter dataset was straightforward. One of the first steps was to determine if any duplication of the data existed within the set. This was achieved by using the ID column since each Kickstarter is assigned a unique identifier. If any duplication did exist, it could be assessed and removed. Fortunately,

no duplication entries were found. The ID category was then dropped since it would not provide any further contributions to our analysis.

The remaining columns were then analyzed to remove irrelevant or redundant data. **Pledged**, **usd_pledged** and **goa**l, were eliminated since their values were already reflected in **usd_pledged_real** and **us_goal_real attributes.**

Next, we looked at the different states each kickstarter reached. According to the data, each particular campaign wound up in one of six states: failed, cancelled, successful, live, undefined, and suspended. Figure 1 below shows the total number of kickstarters tallied for each state.

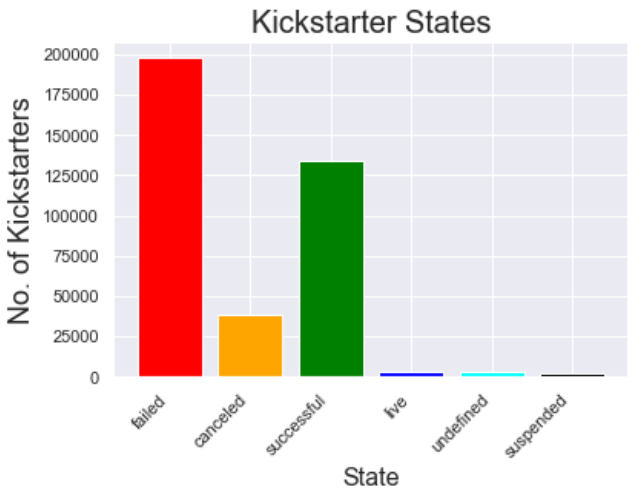

Figure 1.

A failed state indicates a kickstarter was unable to achieve its funding goal. A cancelled state indicates that the individual who launched the project, for one reason or another, aborted the campaign before the kickstarter ended. A live state indicates that the project was still in its funding round when the data was collected. The undefined state is for kickstarters whose campaigns have yet to launch, and suspended are campaigns which were stopped by the kickstarter platform.

Since the goal of the project is to predict success or failure of a kickstarter campaign, any columns that did not explicitly meet these two states were dropped.

Figure 2.

The above figure illustrates the remaining data set after further filtering. Interestingly, almost 60% of the kickstarters which launch fail to achieve their target goals while only about 40% find the necessary financial support from their backers to launch their project.

With the preprocessing and data cleaning completed, the next step was to perform exploratory data analysis to get better insights into the dataset, to identify important attributes, transform the data into meaningful and useful representations, further detect and identify outliers/anomalies, and to begin to piece together the components that would be necessary for our logistic regression model.

8.2 Exploratory Data Analysis

Our Exploratory Data Analysis revealed some interesting differences between failed and successful projects. First we plotted the total count of projects launched, categorized by the Main Category attribute.

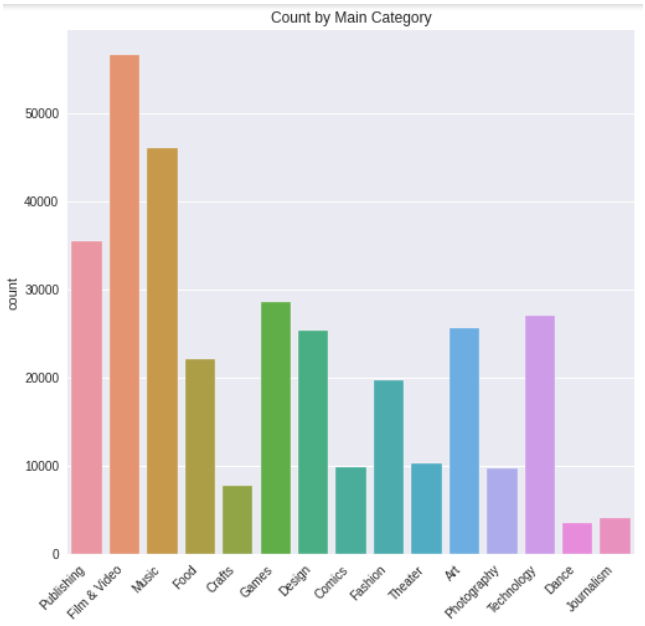

Figure 3.

In Figure 3, we can see that Film & Video projects are the most common, and Dance projects are the least common.

We also looked at the Main Category counts broken down by successful projects and failure projects, which provided more interesting details:

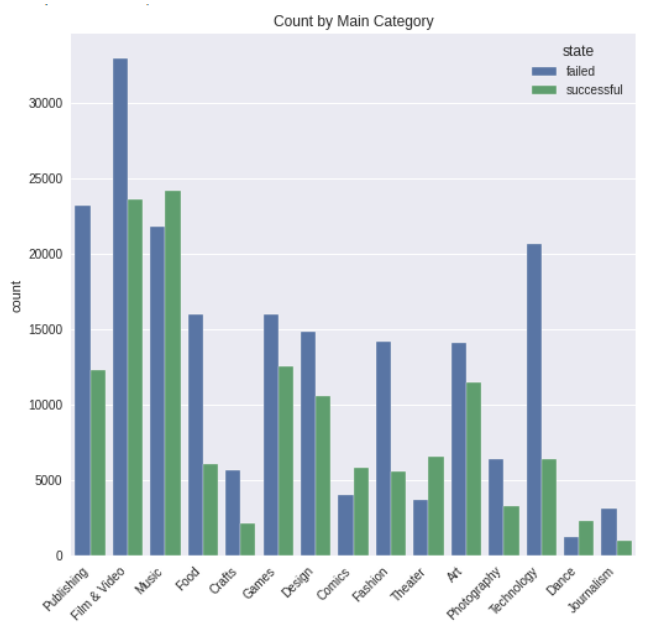

Figure 4.

The bar chart presented in Figure 4 was more interesting because it allowed us to see which Main Category was more likely to succeed. From this chart, we can see that the only categories where success is more common than failure are Music, Comics, Theater and Dance. Although Dance has the fewest number of projects launched, it does have the highest success rate. In fact, it has over a 65% success rate. In contrast, Technology, which we can see from Figure 4 has the lowest success rate, has only a 24% success rate.

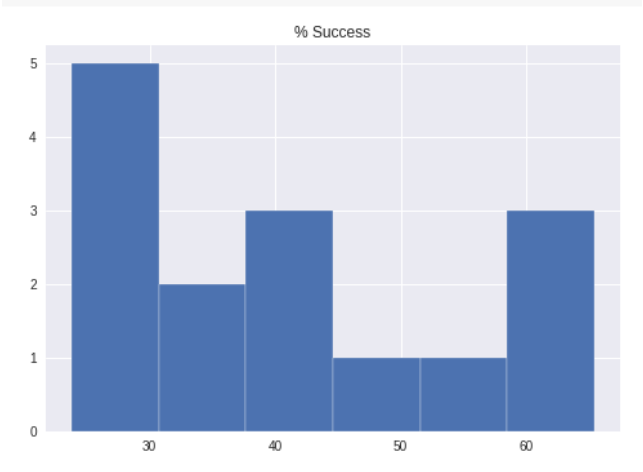

Figure 5.

As previously stated, only four main categories have a success rate above 50%. Figure 5 shows that of the remaining 11 projects, 5 of them have success rates under 30%.

To further analyze success and failure rates of projects we looked at their financial aspects. Projects with lower goals tended to have higher success rates than those with higher goals. As shown in Figure 6 below, the median goal for successful projects was slightly less than $4,000 while the median goal for failed projects was close to $7,500. We further noted that successful projects had more backers than unsuccessful projects, but that the average contribution per backer was relatively consistent regardless of whether the project was successful or failed.

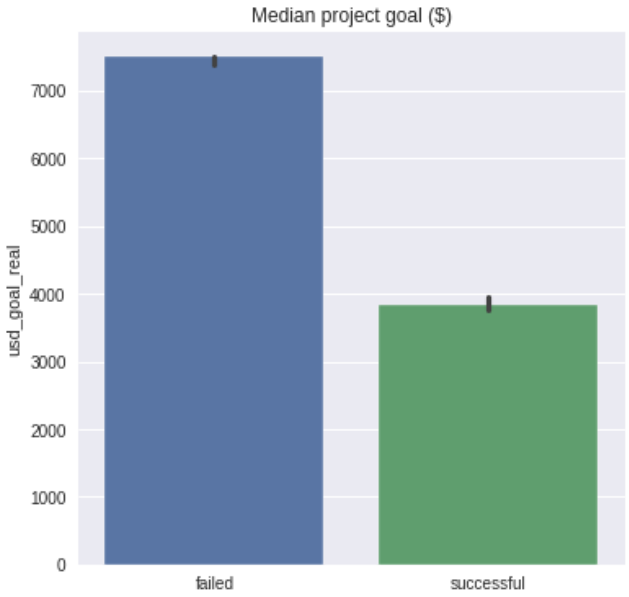

Figure 6.

REFERENCES

[1] Daniel Kupka, 2019. “We Analyzed 331,000 Kickstarter Projects. Here’s What We Learned About Kickstarter Success”. *https://medium.com/@daniel.kupka/we-analyzed-331-000-kickstarter-projects-here-s-what-we-learned-about-crowdfunding-success-63b341b025ac*

[2] Justin S Gage, 2017. “Kickstarter Analysis and Prediction”. *https://www.kaggle.com/gagejustins/kickstarter-analysis-and-prediction*

[3] Laura Lewis, 2019 “Using Machine Learning To Predict Kickstarter Success”,. https://towardsdatascience.com/using-machine-learning-to-predict-Kickstarter-success-e371ab56a743