Analyzing Kickstarter Campaign Success Through Web-Scraped Data

1st Jaden Earl
dept. Statistics
Brigham Young Univeristy
Provo, USA

***Abstract*—This paper presents an analysis of over 10,000 Kickstarter campaigns collected through web scraping using Beautiful Soup. The analysis focused on three key metrics for each campaign: the number of backers, total funds raised, and goal amount. By deriving additional metrics such as average pledge per backer and performing business-oriented calculations including unit cost, overhead cost, profitability, and break-even analysis, I provide a novel perspective on the financial dynamics of crowdfunding. A multi-linear regression model predicting the log-percent of the funding goal achieved explained 45% of the variance in campaign success. These findings offer insights for entrepreneurs evaluating the feasibility of launching successful Kickstarter campaigns.**

***Index Terms*—crowdfunding, kickstarter, web scraping, regression analysis, business analytics**

I. Introduction

Crowdfunding has emerged as a popular approach for entrepreneurs and creators to secure funding for innovative ideas. Among the many platforms, Kickstarter has stood out as a leading choice, hosting thousands of campaigns annually. Understanding the factors contributing to campaign success is crucial for both campaign creators and prospective backers. Previous studies, such as those by Mora-Cruz and Palos-Sanchez (2023) [1] and Talukder and Lakner (2023) [2], have explored the broader dynamics of crowdfunding platforms and their role in social entrepreneurship.

This study analyzes over 10,000 Kickstarter campaigns from the past five years, leveraging data obtained through web scraping. Using Python's Beautiful Soup library, we extracted publicly available information in accordance with Kickstarter's robots.txt file, which permits scraping except for specific private or administrative endpoints.

The collected data includes the number of backers, total funds raised, and campaign goal amounts. From these metrics, we derived additional insights such as the average pledge per backer. Furthermore, we conducted a series of business-oriented calculations, including the estimation of unit costs, overhead costs, and profitability. To explore the factors influencing campaign success, we employed a multi-linear regression model to predict the log-percent of the funding goal achieved, which accounted for 45% of the observed variance.

The implications of this analysis are significant for individuals and organizations considering Kickstarter as a platform for funding. By understanding the interplay of costs, pledges, and campaign goals, creators can better assess the viability and risks of their campaigns.

II. Data Cleaning and Wrangling

To analyze Kickstarter campaign performance, raw data obtained via web scraping was systematically cleaned and transformed. The cleaning process ensured accuracy and usability for downstream analysis.

A. Data Preprocessing

The raw dataset contained information such as campaign goals, total funds raised (in USD), and the number of backers. Specific steps included:

- Converting critical variables (e.g., USD pledged, goal) into numeric formats to allow mathematical operations.
- Ensuring unique campaign entries by deduplicating based on the campaign slug.
- Filtering campaigns with fewer than 5 backers, as they provide insufficient data for meaningful analysis.

From these preprocessed variables, additional metrics were derived:

- **Goal in USD** ($goal\_usd$): Converted campaign goals into USD using the exchange rate.

$$\text{goal\_usd} = \text{goal} \times \text{usd\_exchange\_rate} \tag{1}$$

- **Surplus** ($surplus\_usd$): The difference between total funds pledged and the campaign goal.

$$\text{surplus\_usd} = \text{usd\_pledged} - \text{goal\_usd} \tag{2}$$

- **Average Pledge** ($averagePledge$): The average contribution per backer.

$$\text{averagePledge} = \frac{\text{usd\_pledged}}{\text{backers\_count}} \tag{3}$$

- **Percent of Goal Reached** ($percent\_of\_goal\_reached$): The ratio of funds pledged to the campaign goal.

$$\text{percent\_of\_goal\_reached} = \frac{\text{usd\_pledged}}{\text{goal\_usd}} \tag{4}$$

B. Business Metrics

I introduced business-related metrics to assess the financial viability of Kickstarter campaigns. These metrics were inspired by concepts such as break-even analysis and cost structure.

1) Unit Cost and Variable Cost ($vCost$): Given an assumed unit cost as a percentage of the average pledge, the variable cost per unit was calculated as:

$$vCost = \text{averagePledge} \times \text{unitCostPercentage} \tag{5}$$

This represents the per-unit cost of fulfilling a campaign reward.

2) Goal Value ($goal\_value$): The goal value accounts for overhead costs relative to the break-even percentage of the funding target.

$$goal\_value = \frac{\text{overhead}}{\text{breakEvenPercentage} \times (1 - \text{unitCostPercentage})} \tag{6}$$

This reflects the funding required to cover fixed costs and achieve the break-even target.

3) Break-Even Quantity ($breakQ$): The break-even quantity estimates how many pledges are needed to cover the campaign goal at the assumed break-even percentage:

$$breakQ = \frac{\text{goal\_usd}}{\text{averagePledge}} \times \text{breakEvenPercentage} \tag{7}$$

4) Profitability ($profit$): The potential profit was computed assuming successful funding:

$$\text{profit} = (1 - \text{breakEvenPercentage}) \times \text{goal\_value} \times (1 - \text{unitCostPercentage}) \tag{8}$$

This captures the residual funds after meeting costs and achieving the break-even target.

C. Logical Justification

These variables make sense within a business and crowdfunding context for several reasons:

- *Percent of Goal Reached* provides a normalized measure of campaign success.
- *Variable Costs* ($vCost$) and *Overhead* quantify how operational costs scale with production.
- *Break-Even Analysis* helps creators understand the funding and pledges needed to reach profitability.
- *Profit Metrics* allow campaigns to estimate potential returns and assess financial viability.

Fundamentally, I are uncertain about the actual break-even quantity, fixed costs, and profitability for these Kickstarter campaigns. However, our model allows users to set these parameters for their own campaigns. As such, I assume that all campaigns operate with the same relative cost structure, enabling a generalized analysis.

Together, these metrics offer a framework for evaluating the feasibility and financial sustainability of Kickstarter campaigns.

III. Multi-Linear Regression Model

To understand the factors that influence the percentage of a campaign's goal achieved, I employed a multi-linear regression model. The response variable was the logarithm of the percent of goal reached ($\log_{10}$(percent_of_goal_reached)), and the predictors included the logarithm of the average pledge, the logarithm of the campaign goal in USD, and interaction terms with the campaign's parent category:

$$\begin{aligned}\log_{10}(\text{percent\_of\_goal\_reached}) \sim{}& \log_{10}(\text{averagePledge}) \\ &\times \log_{10}(\text{goal\_usd}) \\ &\times \text{parent\_category}.\end{aligned} \tag{9}$$

The inclusion of interaction terms reflects the hypothesis that relationships between campaign success, goal size, and average pledge may vary significantly across campaign categories.

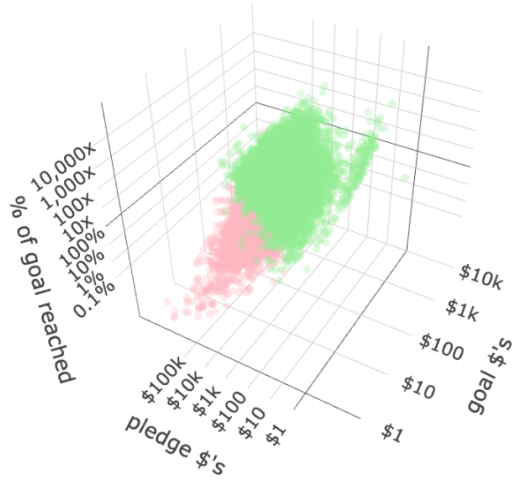

Fig. 1. Log Log Plot that Represents the Model, Green Points Being Successful Campaigns and Pink Points Being Campaigns That Did Not Reach Their Goal

A. Model Exploration

Several alternative models were explored, including:

- Logistic regression on whether a campaign was successful (binary outcome).
- Regression on the absolute distance between funds received and the campaign goal.
- Box-Cox transformations to stabilize variance and improve model fit.

However, these approaches failed to capture the underlying patterns due to extreme non-linearity in the data, making the multi-linear model the most robust choice.

B. Model Performance

The multi-linear regression model explained approximately 40% of the variance in the logarithm of the percentage of a

campaign's funding goal achieved ($R^2 = 0.4083$; Adjusted $R^2 = 0.4055$). This performance reflects a strong relationship between predictors such as the average pledge amount, the campaign goal in USD, and campaign categories. The model was statistically significant overall ($p < 2.2 \times 10^{-16}$).

Key coefficients provide insights into the relationships:

- The logarithm of the average pledge had a positive effect, indicating that higher average pledge amounts are associated with a higher percentage of goal completion.
- The logarithm of the campaign goal had a significant negative impact, meaning larger goals tend to have lower percentages of completion.
- Certain campaign categories (e.g., Games, Photography) significantly interacted with predictors, reflecting their unique effects on campaign outcomes.

While the model captures key trends, higher-order interactions and the complexity of crowdfunding campaigns suggest remaining variance and non-linear patterns.

C. Model Diagnostics and Validation

Model diagnostics were performed to evaluate the assumptions of the regression model, including residual analysis, multicollinearity checks, and influence diagnostics.

1) Residual Analysis: The residuals were assessed using QQ plots and histograms to check for normality:

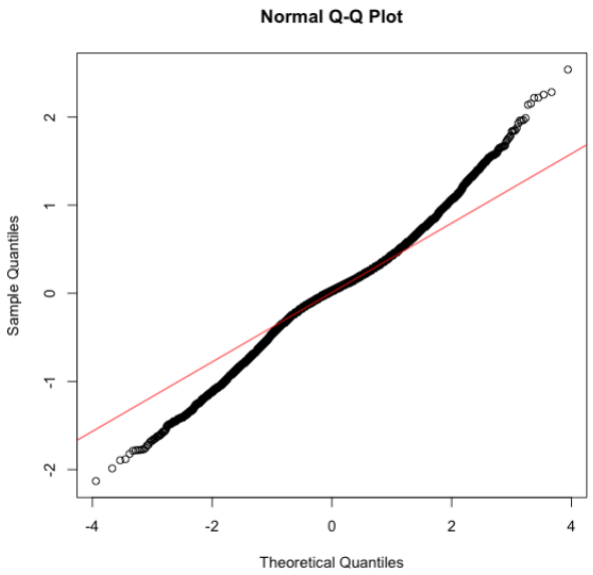

Fig. 2. QQ-Plot of Model Residuals

2) Heteroscedasticity Check: The Breusch-Pagan test was performed to assess heteroscedasticity. The test indicated significant heteroscedasticity ($p < 2.2 \times 10^{-16}$), suggesting that variance is not constant across predictors.

3) Influence Diagnostics: Cook's distance was analyzed to identify influential points. A threshold of $4/n$ was used to flag data points with high influence:

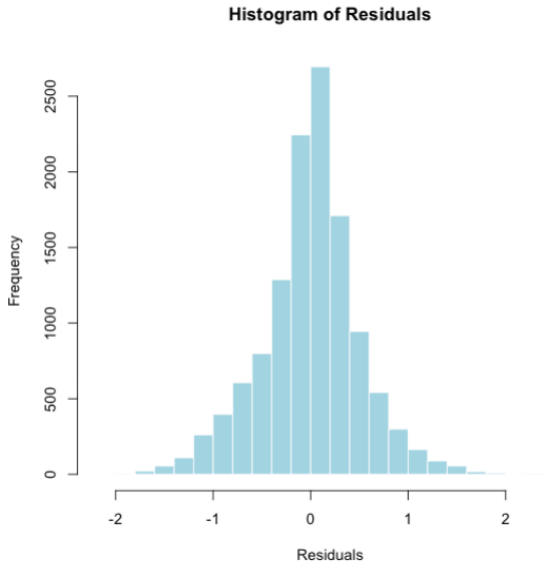

Fig. 3. Histogram of Residuals

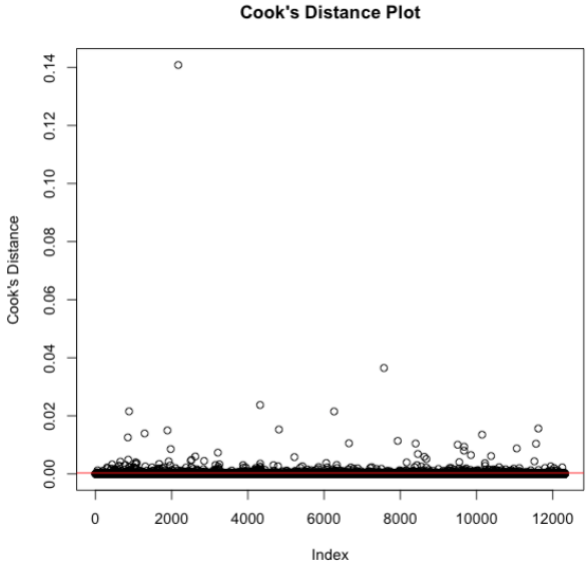

Fig. 4. Cook's Distance Plot

4) Multicollinearity: Variance Inflation Factors (VIF) were calculated to detect multicollinearity among predictors. High VIF values for interaction terms highlight challenges in model complexity and interpretability.

D. Model Performance

The Akaike Information Criterion (AIC) for the selected model was 18254.34, which supports its relative fit among other candidates.

E. Conclusion

While the model highlights important trends, significant non-linearity and heteroscedasticity remain challenges. Future work may focus on non-linear methods or machine learning approaches to capture more complex relationships.

IV. Toy Example: Predicted Kickstarter Outcome

To illustrate the application of the model, consider a hypothetical Kickstarter campaign in the Technology category with the following parameters:

- Average Pledge Value: $30
- Campaign Goal: $1875

Using the model's predictions:

- The campaign's **pledge percentile** relative to other campaigns: 13.34%
- The campaign's **goal percentile**: 29.61%
- The **predicted probability of meeting the goal**: 44.7%
- The **expected revenue**: $2189.24

These results demonstrate the model's ability to provide actionable insights, such as the likelihood of success and expected revenue. By adjusting key parameters (e.g., pledge amounts, goal size), campaign creators can evaluate scenarios to optimize outcomes and assess financial feasibility.

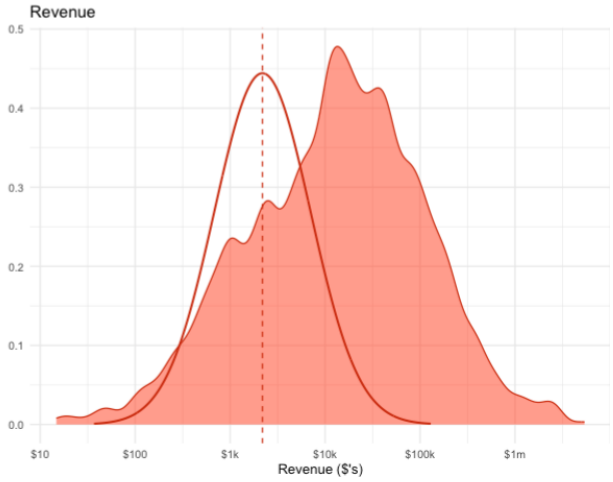

Fig. 5. Predicted Revenue Probability Density (line) for Hypothetical Campaign with Density of all Campaigns Revenue (dark red area)

V. Conclusion

This analysis of Kickstarter campaigns provides several key insights into the factors influencing crowdfunding success:

- Higher average pledges and smaller campaign goals are associated with a greater percentage of goal completion.
- Campaign categories exhibit distinct interactions with key predictors, highlighting the importance of industry-specific strategies.
- The model captures 40% of the variance in campaign success, providing a strong baseline for predictive analysis in crowdfunding.

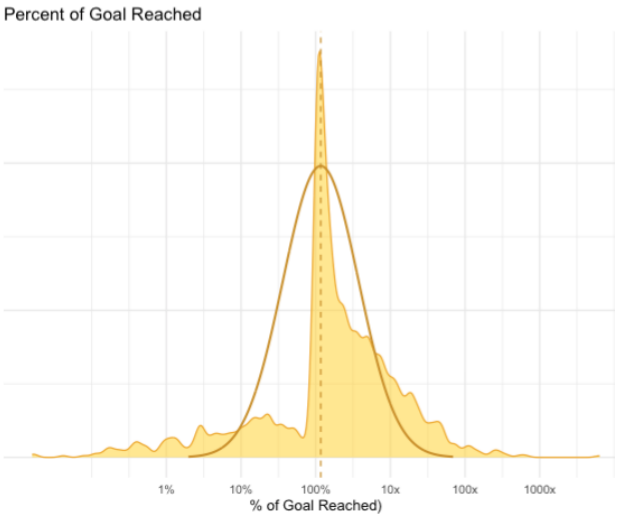

Fig. 6. Predicted Percentage Completion of Goal Probability Density (line) for Hypothetical Campaign with Density of all Campaigns Percentage of Goal Achieved (yellow area)

The unique contribution of this work lies in its integration of business metrics, such as unit costs, overhead, and break-even analysis, with campaign success predictions. By allowing entrepreneurs to set their parameters, the analysis provides tailored insights into financial feasibility and profitability.

This research can inform Kickstarter creators on how to set realistic funding goals and optimize average pledge sizes to improve their chances of success. It also offers a broader framework for evaluating entrepreneurship decisions on crowdfunding platforms.

However, there are limitations. The model assumes linear relationships between predictors and outcomes, which may not fully capture the complex, non-linear nature of crowdfunding success. Additionally, the analysis is constrained to publicly available data, which excludes critical factors like marketing efforts or product quality.

Future research should explore advanced non-linear models, such as machine learning approaches, to improve predictive accuracy. Incorporating external data, such as campaign descriptions, social media metrics, or backer engagement, could further enhance the model's explanatory power.

In conclusion, this study offers a robust and actionable framework for understanding crowdfunding dynamics, with clear implications for Kickstarter creators and the broader entrepreneurial community. By leveraging these insights, campaign creators can better navigate the challenges of crowdfunding and increase their likelihood of success.

References

[1] A. Mora-Cruz and P. R. Palos-Sanchez, "Crowdfunding platforms: a systematic literature review and a bibliometric analysis," *International Entrepreneurship and Management Journal*, vol. 19, no. 3, pp. 1–26, 2023.

[2] S. C. Talukder and Z. Lakner, "Exploring the Landscape of Social Entrepreneurship and Crowdfunding: A Bibliometric Analysis," *Sustainability*, vol. 15, no. 12, pp. 1–226, 2023.