# Potential Write-Up Format

## **1. Introduction**

Kickstarter is a crowd-sourcing platform that allows creators of all sorts share their prospective work to attract community funding. It launched in 2009 and has since become a Public Benefit Corporation that has funded more than 270,000 projects and raised more than $8 billion dollars. Kickstarter creators post their ideas or projects on the website and share it with their friends and supporters, who, in turn, share the project in their networks, and so on. More than 24 million people from all over the world have helped fund Kickstarter campaigns. Project cover a wide range of categories including art, publishing, design, and technology. “Kickstarter lifts the creative class, gives people the tools to pursue daring ideas on their own terms, and helps creators build communities around their work.” (www.kickstarter.com)

* **SMART Questions**
  + *Your research questions, and how did they come up?*

A head with gears in the brain

Description automatically generated

For this project, our research questions were designed using the **SMART** framework—Specific, Measurable, Achievable, Relevant, and Time-bound. This analysis seeks to answer the following questions:

1. Which variables most influence success or failure?
2. Can a logistic regression model accurately predict the success or failure of a Kickstarter campaign?
3. Can a Decision Tree accurately predict the success or failure of a Kickstarter campaign?
4. What are the top five (5) categories with the highest number of successes?
5. What percentage of all campaigns were successful compared to failed?

Through these questions, we aim to determine if Kickstarter campaign outcomes are predictable and identify which features/attributes contribute to the success of a campaign.

## **2. Dataset Description & Preparation**

**Dataset Overview**

The dataset was sourced from Kaggle and contains 378,661 observations with 15 variables. These variables include:

* **backers:** The total number of backers who supported a project.
* **currency:** The currency in which the project was originally launched.
* **country:** The country from which the project was launched.
* **main\_category:** The primary category of the project (e.g., Music, Technology).
* **state:** The final status of the project, indicating whether it was successful, failed, or canceled.
* **usd\_pledged\_real:** The total amount of money pledged to a project in USD.
* **usd\_goal\_real:** The funding goal set by the project creators in USD.

**Data Preprocessing/Cleaning**

* Mention removal of variables or rows
* Any missing /NA?
* Any data type conversions?

**\*\* Data Limitations *(if any)* – Can also go at the end? (see above “Conclusions”)**

## **3. Literature Review (if applicable)**

- Any previous research/ analysis on this?

## **4. Exploratory Data Analysis**

- Descriptive Statistics

- Plots and analysis

**A red and green graph

Description automatically generated**

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

|  |  |
| --- | --- |
| **Top 5 Categories with the Highest Number of Successful Projects:** | |
| **Main Category** | **Count** |
| Music | 24197 |
| Film & Video | 23623 |
| Games | 12518 |
| Publishing | 12300 |
| Art | 11510 |

Identifying the top five categories with the most successful projects, the table above highlights the areas where Kickstarter campaigns tend to achieve the most success. These categories are Music, Film & Video, Games, Publishing, and Art, with Music accounting for the greatest number of successful projects with a value of 24,197, followed closely by Film and Video with 23,623 successful projects.

**A graph of a bar chart

Description automatically generated with medium confidence**To make these insights more visually appealing and intuitive, these values can be represented in a bar chart as seen below:

## **A graph of a number of backers Description automatically generated**

## **5. Modeling Techniques & Evaluation**

- *How did you select and determine the correct model to answer your question?*

*- Build Models*

*- Analyze/ Evaluate Models using evaluation metrics?*

**Logistic Regression**

A screenshot of a computer screen

Description automatically generated

This logistic regression model uses three predictors, **backers**, **usd\_pledged\_real**, and **main\_category,** to predict the success of Kickstarter projects. The dataset was split into a training set (70%) and a testing set (30%) using the ***train\_test\_split()*** function. The model’s accuracy can be seen below:

* **Test set accuracy:** 82.41%
* **Train set accuracy:** 82.16%

The similarity in accuracy for both test and train sets indicates that the model generalizes well to unseen data.

From this model it is observed that:

* **Backers (Coefficient = 0.024616; Odds Ratio = 1.0249):**
  + Each unit increase in backers (i.e. each additional backer), the **odds of success** increases by a factor of approximately 1.03.
* **USD Pledged (Coefficient = 0.000042; Odds Ratio = 1.00004):**
  + Although the effect of this feature is small, a 1 unit increase in pledged funds has a positive effect on success likelihood, as it increases the odds of success by a factor of 1.
* **Main Categories:**
  + This model uses the “Art” category as the baseline for comparison. Some insights are as follows:
    - Categories such as **Comics, Dance, Music, Film and Video and Theater,** have positive coefficients and odds ratios greater than 0, meaning that campaigns in these categories are more likely to succeed compared to campaigns in the **Art** category. For example, campaigns in the **Theater** category increases the odds of success by a factor of 2.89 when compared to those in the **Art** category.
    - Categories like **Crafts, Fashion, Technology, Games, and Food** have negative coefficients, indicating that campaigns in these categories have a lower odds of success when compared to the baseline category, **Art**. For example, projects considered within the **Games** category are less likely to be successful than those in the **Art** category as it odds ratio is approximately 0.3397.

A colorful squares with numbers

Description automatically generated

**From the confusion matrix, classification metrics can be derived and evaluated.**

A screenshot of a computer screen

Description automatically generated

**F1-score:**

* The model’s F1-Score of 0.82 (82%)represents the balance between precision and recall, indicating that the model is performing well overall. However, there is room for improvement in minimizing false negatives.

White text on a black background

Description automatically generated

**False Positive Rate (FPR): 5.89%**

* This FPR value indicates that among projects that 5.89% of failed projects were incorrectly classified as successful by the model.

**False Negative Rate (FNR): 34.99%**

* The FNR of 34.99% implies that 34.99% of successful projects were incorrectly classified as not successful *(failed)* by the model.

This high FNR indicates the model has difficulty identifying successful projects. To address this high FNR value, the cutoff was lowered to 0.3. After refitting this model with new threshold, the following metrics were obtained:

A screenshot of a computer

Description automatically generated

**Accuracy**: **83%**

– The model correctly predicted the outcome (success or failure) 83% of the time across all the projects.

**Pecision: 84%**

– Although the average precision is approximately 84%, there is a large difference between the precision of the two classes. The model has a higher precision at predicting failures than successes.

**F-1 Score: 84%**

* While slightly lower for successes than for failures, the model still does generally well with balancing precision and recall, with a slight increase at the new threshold of 0.3 compared to 0.5. This indicates the model is good for capturing successful projects.

With further evaluation, the confusion matrix of this model at a threshold of 0.3 can be seen below:

A chart with numbers and a few squares

Description automatically generated with medium confidence

The new FPR and FNR is calculated from this matrix.

A black background with white text

Description automatically generated

Compared to the original model with the default threshold of 0.5, reducing the threshold value has decreased the FNR and increased the FPR.

**False Positive Rate (FPR):**

* The FPR is calculated to be **18.64%,** indicating that 18.64% of the failed projects were incorrectly classified as successful.

**False Negative Rate (FNR):**

* The FNR value of **13.55%** shows that 13.55% of the actual positive samples (i.e. successful projects) were incorrectly classified as failed. Although the FPR has increased, there is more of a balance between the two rates.

When comparing the models at different cutoffs, overall, lowering the threshold from 0.5 to 0.3 resulted in a slight improvement in accuracy from (82.41% to 83%). The F-1 score increased slightly at a cutoff of 0.3, indicating a better balance between the precision and recall.

## **6.Discussion**

*- Interpret results*

*- What Predictions Can You Make from Your Models? Examples?*

*- How good is your model?*

- *How reliable are your results?*

\*\* LIMITATIONS:

*- What additional information or analysis might improve your model results or work to control limitations?*

## **7. Conclusion.**

- Summary

- How Do These Answer the Research Questions?

## **8. References (APA Style)**