# 

# **Predicting Kickstarter Campaigns**

**Fardin Hafiz, Leshauna Hartman**, **Rachel Thomas**, **Tanya Visser**

# **The George Washington University**

# **Introduction to Data Mining**

# **Professor Divya Pandove Narula**

# **Tuesday 17th December 2024**

## **Introduction**

Kickstarter is a crowd-sourcing platform that allows creators of all sorts to share their prospective work to attract community funding. Launched in 2009, Kickstarter 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 them with 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. Projects 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](http://www.kickstarter.com)).

We are interested in predicting the success or failure of Kickstarter projects and identifying the characteristics of a project that most influence this outcome. This has practical implications as creators can tweak their projects to reflect a higher chance of success. Additionally, backers can make better decisions about which projects to fund. This saves time, money, and resources for both the creators and the backers.

Kickstarter is an “all or none” funding scheme, meaning if a campaign does not raise the entirety of its goal funding, it gets zero funding, all the money pledged everts to the donors, and the project fails.

## **Research Questions**

Our research questions were designed using the **SMART** framework – Specific, Measurable, Achievable, Relevant, and Time-bound. This analysis seeks to answer the following:

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 percentage 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.

## **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**

After sourcing the dataset, it was filtered to include only campaigns that were successful or failed – rows reflecting any other campaign states were removed. This reduced the dataset to approximately 331,000 observations. A new variable called Duration, was created by finding the difference between the `launched` and `deadline` variables. `main\_category`, `currency`, `state` and `country` variable were converted to categorical data types. A subset called kickstarter\_final was created to include `main\_category`, `currency`, `state`, `backers`, `country`, `usd\_pledged\_real`, `usd\_goal\_real`, and `Duration`.

**Literature Review**

Online crowdfunding platforms are an increasingly popular way for ordinary people to finance a wide variety of projects ranging from creative arts to healthcare support. Though there have been many platforms, Kickstarter is regarded as the largest and most impactful. In a 2016 study, Ethan Mollick of the Wharton School at the University of Pennsylvania, reported that each dollar given to projects via Kickstarter resulted in a mean of $2.46 in additional revenue (though this was not spread evening though categories). He also reported that Kickstarter projects had resulted in more than 5,000 ongoing full time jobs besides those of the creators, and more than 160,000 temporary positions. The successful campaign also resulted in more than 2,600 patent applications (Mollick, 2016).

However, as more campaign have been launched, there has been an observed decrease in success rate, suspected to be due to campaign launches without sufficient preparation or experience. Tran, et al, showed that campaigns with significantly lower goals and significantly increased advertisement (via Twitter posts), were more successful (Tran et al., 2016). The role of advertisement is tangential, but essential to acquiring backers. Carradini & Fleishmann (2023) found images, links, and the presence of a project video significantly contributed to success, while gifs and galleries did not. Additionally, Tian (2021) found whether a campaign was selected to be highlighted as an editor’s pick significantly impacted success. Editor’s picks are featured on Kickstarter’s homepage and thus receive more overt publicity than those not featured.

**Exploratory Data Analysis**

We started the analysis with some exploratory graphing, to understand the variables better.

**A red and green graph

Description automatically generated**

Figure 1. Distribution of final campaign state

You can see above (Fig. 1) that failed projects exceed successful projects across this dataset. This is not surprising, given the previous body of knowledge about crowdfunding in general, and Kickstarter specifically.

Next we examined the distribution of project outcomes by currency and country. You can clearly see in the graphs below that projects based in the US and funded with the US Dollar far outnumber those in any other currency (Fig. 2) or country (Fig. 3).

A screenshot of a graph

Description automatically generated

Figure 2. Distribution of final campaign state by country

A screenshot of a graph

Description automatically generated

Figure 3. Distribution of final campaign outcome by currency

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 Dance, Theater, Comics, Music, and Art.

|  |  |
| --- | --- |
| **Top 5 Categories with the Highest Percentage of Successful Projects:** | |
| **Main Category** | **Percentage** |
| Dance | 65.44% |
| Theater | 63.8% |
| Comics | 59.14% |
| Music | 52.66% |
| Art | 44.89% |

To make these insights more visually appealing and intuitive, these values can be represented in a bar chart as seen below (Fig. 4):

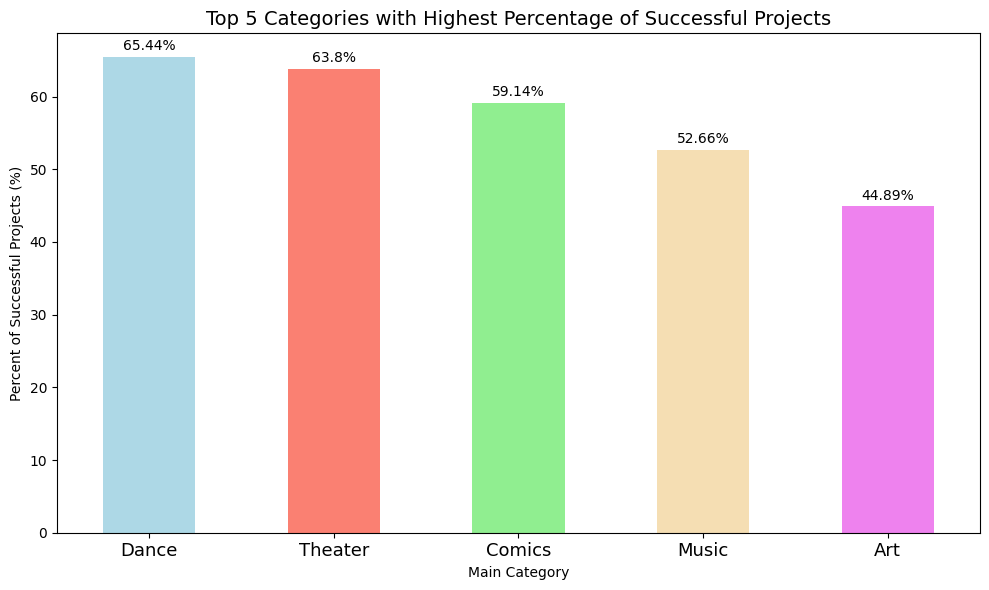
****

Figure 4. Top five categories with highest percentage of successful projects.

**A graph of a number of backers

Description automatically generated**

Successful campaigns have more backers and smaller funding goals, while failed campaigns have fewer backer and large to extremely large finding goals (Fig. 5). This is intuitively logical as large goals may dissuade backers due to a perception of unattainability leading to failure.

Figure 5. Final campaign state by number of backers and funding goal

## 

Figure 6. Success rate by main category and median funding goal in USD for all countries.

For all countries, the categories with the highest success rates had lower funding goals and the categories with the lowest success rates had the highest funding goals (Fig. 6). It appears that most technology campaigns fail related to their very high funding goals.

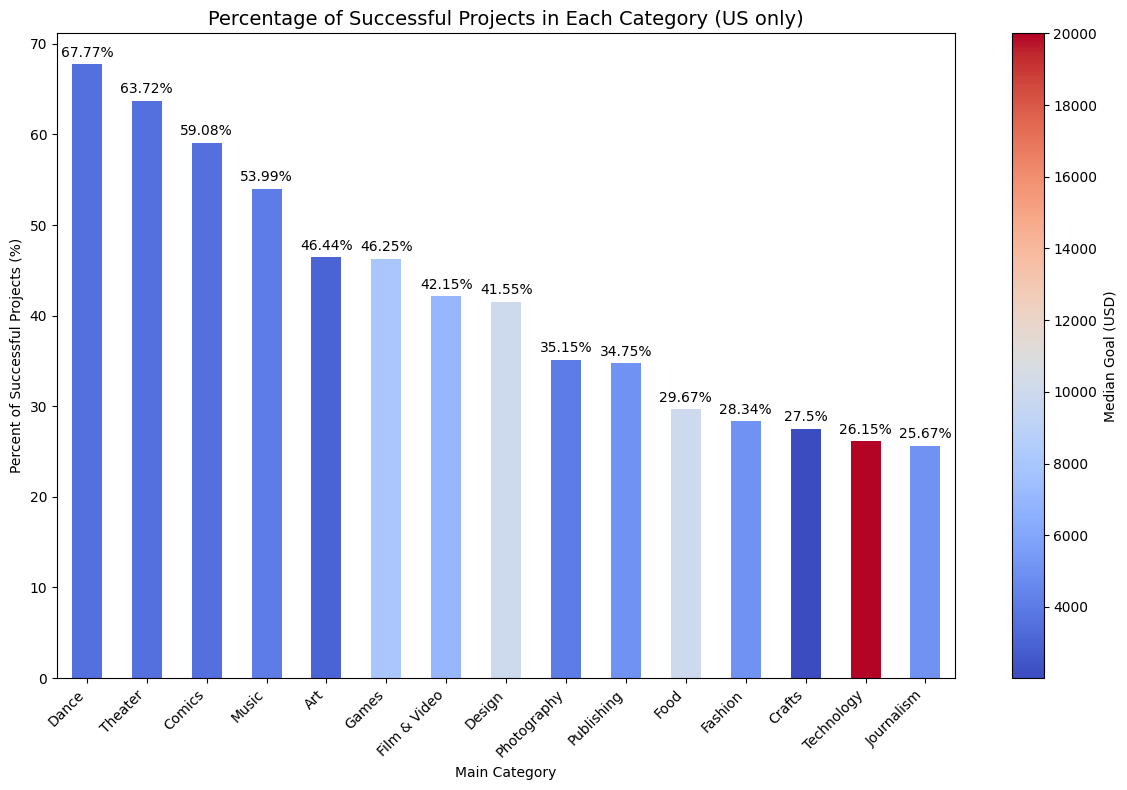


Figure 7. Success rate by main category and median funding goal in USD for US projects only.

When we subset the data to include US projects only, we find the same result: the categories with the highest success rates had lower funding goals and the categories with the lowest success rates had the highest funding goals (Fig. 7). However, technology appears to slightly improve their success rate despite having the highest funding goals.

A green and white graph

Description automatically generated

*Figure 8. Percentage of the goal met between failed and successful campaigns*

Figure 8 analyzes the percentage of the asking goal met by the campaign comparing failed and successful campaigns. We can see that failed campaigns meet about 9.06% of their campaign goal while successful ones go over their requested goal on average by a percentage of 855%.

## A green and red pie chart Description automatically generated

*Figure 9. Pie chart depicting the distribution of projects between zero and 1+ backers*

Figure 9 shows the percentage of campaigns that had zero backers and 1+ backers. About 14.7% of campaigns did not get off the ground and had zero backers while 85.3% of projects had at least one person backing.

A graph of a number of backers

Description automatically generated

*Figure 10. Histogram of backers with Outliers Removed*

Figure 10 depicts a histogram of backers with the outliers removed. The outliers were removed to get a good visualization of the graph. The graph shows that most common number of backers is between 0 to 20.

## **Modeling Techniques & Evaluation**

This is a binary classification problem, predicting the likelihood of a successful versus failed Kickstarter campaign. We used both a decision tree (a machine learning model), and a logistic regression model to determine how accurately we could predict success.

**Classification Decision tree**

To create the decision tree, we created a training set and then fit a tree to max depth 8. This produced a complex tree with 34 leaf nodes, but training error rate of 0. The test error rate for this tree was 5.2% which is good, however the size of the tree makes it very complex to understand and we strongly suspected this tree was overfitting the data.

A diagram of a decision tree

Description automatically generated

Figure 11. Decision tree, max depth 8

We then tried trees at depths 3. 4. and 5. To compare the models at the different depths, we completed cross validation at each and compared the scores using T-tests.

A screenshot of a computer

Description automatically generated

Figure 12. Comparison of cross validation depths to find ideal maximum depth

The model with max depth 4 performed statistically better than the model with max depth 3, but the model with max depth 5 was not statistically better than the model with max depth 4. We also performed a validation curve for the tree which showed the ideal depth is probably around 5-6 based on the point at which both the training score and the validation score are at their highest.

A graph of a graph showing the value of a tree

Description automatically generated

Figure 13. Validation curve for Kickstarter decision tree

However, the models at these depths are very complex and much more difficult to understand. So, we elected to use the model with max depth 4 as we felt the slight increase error rate was acceptable in exchange for a much simpler model with only 15 terminal leaf nodes.

A diagram of a decision tree

Description automatically generated

Figure 14. Final decision tree, max depth 4

To interpret the tree, we generated a text summary which is significantly more readable than the tree plot.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

Figure 15. Text summary of decision tree, max depth 4.

This tree splits first by backers, then by goal, then by pledged amount, and then category, with a small influence of country of backers.

Path 1: <= 12.5 backers, a goal <= $650, pledge amount <= $184.81, and not in Music are likely to fail.

Path 2: <= 12.5 backers, a goal <= $650, pledge amount > #184.81, and not in Publishing are likely to succeed.

Path 3: <= 12.5 backers, a goal > $650, having a specific country, and in Dance will likely fail.

Path 4: >12.5 backers but <= 67.5, a goal <= $4747, and any value pledged is likely to be successful.

Path 5: > 12.5 backers but <= 67.5, a goal is > $4747, and pledged <= $6322.9 are likely to fail, but > $6322.9 are likely to succeed.

Path 6: > 67.5 backers, a goal <= $36970.10, and not Crafts are likely to succeed and if Crafts is likely to fail.

Path 7: >67.5 backers, a goal > $36970, and pledged amount <= $38512.01 are like to fail, but if > $38512.01 are likely to be successful.

Overall, backers are the most significant predictor. Projects with fewer than 12.5 backers are most likely to fail regardless of other factors. Projects with backers between 12.5-67.5 backers increased the likelihood of success as long as goals re small to moderate. Projects with > 67.5 backers have the highest likelihood of success, even with higher funding goals. Small funding goals succeed more often, with projects with a goal less than $4747 (and especially less than $650), are highly likely to succeed assuming they get some backers and some pledged amount. Low pledged amounts lead to failure, especially for high goals. Categories play an overall secondary role, though funding for Craft projects is likely to fail in most scenarios. There may be some small regional effects, but they are not substantial.

We then fit the test data to the model and created a confusion matrix to assess the error.

A blue squares with white text

Description automatically generated

Figure 16. Confusion matrix for decision tree, max depth 4

The tree at depth 3 was reasonably easy to understand and had a training error rate of only 6.9%, and a test error rate of 10.6%. At max depth 4, the training error rate was 6.87% and the test error rate was 8.4%. At max depth 5, the training error was only 1.6% and the test error 6.4%. We decided max depth 4 provided the best balance between accuracy and readability or utility, without risk of overfitting. The decision tree is extremely predictive of Kickstarter campaign outcomes and very clearly highlights the importance of the backers and goal features.

**Logistic Regression**

To further assess the ability to accurately predict the success or failure of a Kickstarter campaign, we built a logistic regression model using three predictors: `**backers`**, `**usd\_pledged\_real`**, and `**main\_category`**. The dataset was divided into a training set (70%) and a testing set (30%) using the ***train\_test\_split()*** function. The model’s summary can be observed below (Fig. 17) and further insights are drawn.

A screenshot of a computer screen

Description automatically generated

Figure 17. Logistic Regression Model Summary

The model achieved an accuracy of 82.41% on the test set and 82.16%, indicating that the model generalizes well to unseen data, as evidenced by similar accuracies for both sets.

From the model’s coefficients, we observed that backers had a positive impact on campaign success. Specifically, for each additional backer, the odds of success increased by a factor of approximately 1.03.While the effect of amount pledged was minimal, it still showed that higher pledge amounts slightly increased the likelihood of success. Regarding main categories, campaigns in categories such as Comics, Music and Theater showed higher odds of success compared to those in the baseline “Art” category. For example, campaigns in the Theater category increased the odds of success by a factor of 2.89 relative to the Art category. Conversely, categories like Fashion, Games and Technology, exhibited lower odds of success, with the Games category showing an odds ratio of approximately 0.3397, meaning that projects in this category are less likely to be successful compared to those in the Art category.

A screenshot of a computer screen

Description automatically generated

A colorful squares with numbers

Description automatically generated  
  
Figure 18a. Classification Report of Logistic Regression Model

Figure 18b. Confusion Matrix of Logistic Regression Model

To assess the model’s performance, we first examined the classification report (Fig. 18a) and confusion matrix (Fig. 18b). The lower precision for failed projects (class 0) compared to successful projects (class 1) indicates that the model occasionally misclassifies successful projects as failed. The higher precision and lower recall for successful projects also suggest that when the model predicts success, it is generally correct, but many successful projects are missed, leading to a lower F1-score of 75% for class 1 compared to 86% class 0. This is reflected in the False Negative Rate (FNR) derived from the confusion matrix, which shows that roughly 35% of successful projects are incorrectly classified as failures. The False Positive Rate (FPR) of 5.89% highlights that fewer failed projects are misclassified as successful.

A graph of a curve

Description automatically generated

Figure 19. ROC Curve of Logit Model

To further evaluate the model, we built the Receiver Operating Characteristic (ROC) curve (Fig. 19), where the Area Under the Curve (AUC) was found to be 90%, indicating that the model performs well with distinguishing between classes. However, to improve performance, we lowered the decision threshold from the default 0.5 to 0.3, which slightly increased accuracy to 83% as seen in the updated classification report below (Fig. 20). This adjustment led to an increase in FPR (18.64%) and a decrease in FNR (13.55%), resulting in a more balanced classification of successful projects.

A screenshot of a computer

Description automatically generated

Figure 20. Classification Report at Threshold 0.3

The new confusion matrix at threshold 0.3 (Fig. 21) provides a more detailed breakdown of the distribution of true positives, true negatives, false positives and false negatives.

*A chart with numbers and a few squares

Description automatically generated with medium confidenceFigure 21. Confusion Matrix at Threshold 0.3*

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

**Logistic Regression – US only**

Since the US had more Kickstarter campaigns in our dataset by far when compared to all the other countries combined, we wanted to see if completing the logistic regression model on data subset only to the US based projects would increase the accuracy of the model and its predictions.

First, we tried a forward step-wise feature selection on 5% of the US data with an 80/20% training/test split and found that `backers`, `usd\_pledged\_real`, `usd\_goal\_real`, `Duration`, and `main\_category` produced the highest accuracy of 0.9984. This almost perfectly predictive result did not seem reasonable and suggested the model was overfitting the training data. So, we tried different variations of these variables and found that anytime `backers` and `usd\_goal\_real` were included in the model, the accuracy was over 0.99. We chose to remove `usd\_goal\_real` since `backers` was chosen to be the most important feature using the forward step-wise feature selection. The VIF created using the features selected with the SFS showed that there was not significant multicollinearity when looking at all of the features together. The results from the SFS and VIF calculation are seen below in Figure 22.

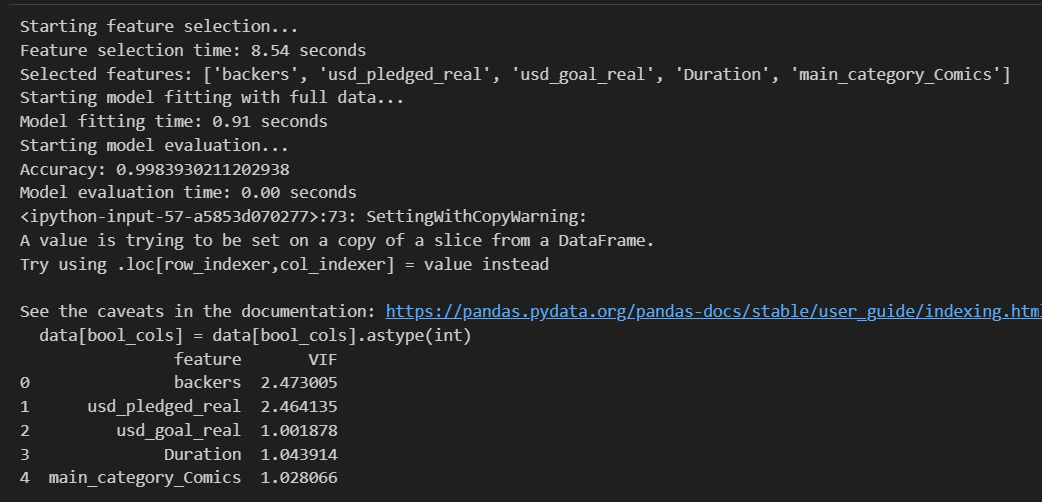


Figure 22. Forward Step-Wise Feature Selection of 5% of the US data and VIF calculation.

The logistic regression model that we created with all of the US data used `backers`, `usd\_pledged\_real`, and the dummy variables for `main\_category` as predictors. We completed an 80/20% training/test split and fit the model. We chose to remove `Duration` because when all of the dummy variables for `main\_category` were used in the model, `Duration`’s VIF increased to above 5 and did not change the accuracy. We can see that all of the coefficients are statistically significant in the model since their p-values are very small. With the remaining 3 variables, the VIF was low for all coefficients (Fig. 24). Results for the logistic regression model fitted with all of the US data is shown below in Figure 23.

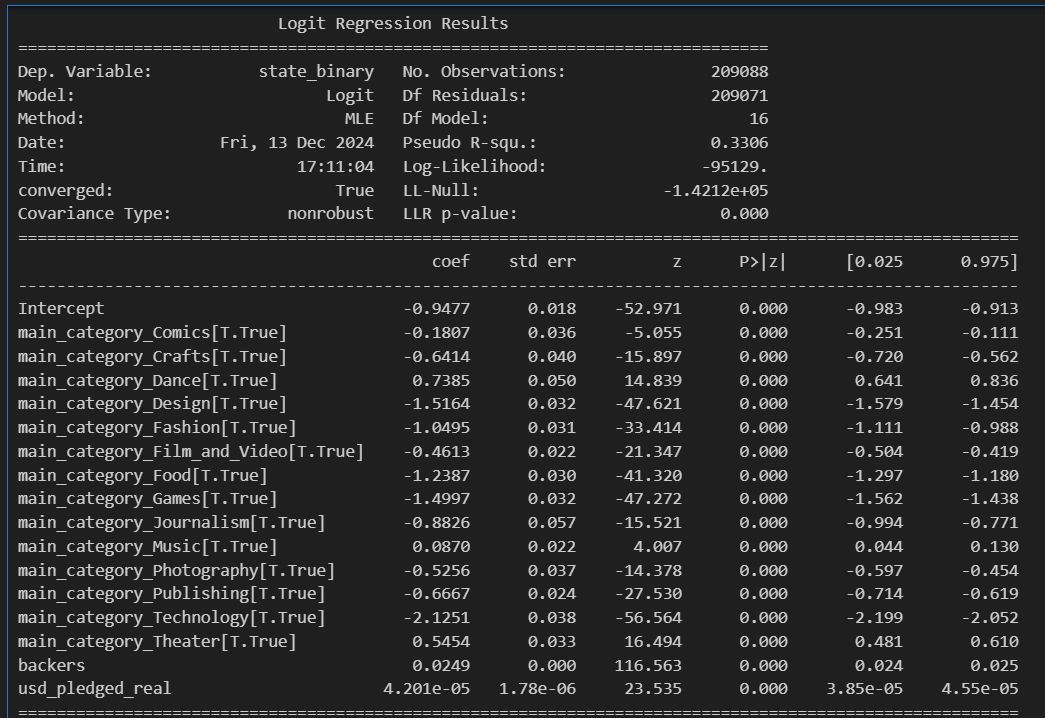


Figure 23. US-only Logistic Regression Model.



Figure 24. US-only Logistic Regression Model VIF calculations.

The test accuracy of the US-only model was 0.8272 (Fig. 25) and the training accuracy of the US-only model was 0.8259 (Fig. 26). This suggests that the model is likely not overfitting. This is a slight increase in accuracy for the US-only model compared to the model with all countries. However, a statistical test would have to be performed in order to decide whether this slight increase is statistically significant.

|  |  |  |
| --- | --- | --- |
|  | All Countries Model | US-only Model |
| Test Set Accuracy | 82.41% | 82.72% |
| Training Set Accuracy | 82.16% | 82.59% |

The classification report was very similar between the testing and training data for the US-only model. When looking at the average F-1 score of 0.81 (Fig. 25), it indicates that the precision and recall for the model overall is well-balanced. However, it is important to note that there is room for improvement regarding the recall for successful campaigns (class 1) which only had a recall of 0.68 compared to the recall for failed campaigns (class 0) of 0.94 (Fig. 25).

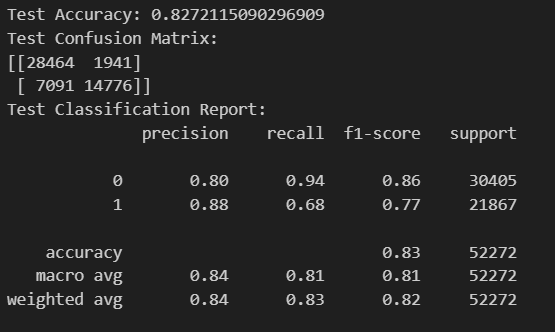


Figure 25. US-only Logistic Regression Model Test Accuracy, Confusion Matrix, and Classification Report.

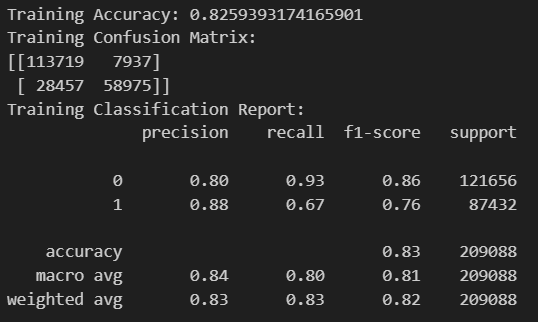


Figure 26. US-only Logistic Regression Model Training Accuracy, Confusion Matrix, and Classification Report.

## 

## **Discussion**

We found that a higher number of backers, a larger pledge amount, a lower funding goal amount, and the specific category that the potential project falls into increase the chance of a successful campaign. This was corroborated by the models that we created which indicated that these four features were the most important for predicting campaign success or failure.

Our models showed high AUC, accuracy, precision, recall, and F1 scores suggesting we created reasonably reliable and predictive models. The decision tree model showed the highest accuracy followed by the logistic regression models, so we would suggest using the nonparametric model to make the most accurate predictions.

Insights from our models can improve the likelihood of success for a creator’s Kickstarter campaign. Especially when reading the decision tree model, creators should be able to figure out reasonable funding goal amounts, number of backers, and pledge amounts for the category that their project falls into. This way they can potentially focus on attracting a certain number of backers with incentives or rewards for helping fund their project. They could also change their projects so that they initially require less funding which would be reflected in their goal; this would potentially get their project off the ground and likely would fund the additional components that they had removed to lower the goal.

**Limitations**

Some potential limitations of this dataset should be discussed in order to see how to improve this project in the future. One limitation is the inconsistency within certain data points. When analyzing the data, we found that the state labeled ‘canceled’ has some potential illogical results. Many ‘canceled’ projects actually met their goal and could be considered ‘successful’. Due to some of these inconsistencies, we did not use the ‘canceled’ state. There were also many data points that were left empty and had to be removed in order to conduct statistical testing. Having missing data could effect any statistical analysis done on the dataset. There were also no variable that include much information each project, besides the genre, which could be important in analyzing why these campaigns succussed or failed.

This leads into our areas of future research, which could be trying to get more information about each campaign. This could be beneficial in building better models to predict success or failure. Another potential area of future research could be adding metrics of how the successful campaigns achieved their set out goals or not. This would be valuable in determining if having a successful Kickstarter campaign actually translates into having success in the goal that was set out. A study like this could also shed light on the importance of having a successful campaign on Kickstarter.

## **Conclusion**

Kickstarter campaigns are highly predictable, with the number of backers and the funding goal as the main drivers of outcome. However, the amount of money pledged is also a strong predictor of success, with successful project raising significantly more money than failed projects. Projects in dance, theater, comics, music, and art are the most successful, while crafts, journalism and technology are the least. Ultimately projects fail more than succeed, but with a smaller initial funding goal and a deliberate plan for attracting backers, it is certainly beyond feasible to achieve success.

## **References**

*About – Kickstarter*. (2024). Kickstarter. <https://www.kickstarter.com/about>

Carradini, S., & Fleischmann, C. (2023). The Effects of Multimodal Elements on Success in

Kickstarter Crowdfunding Campaigns. *Journal of Business and Technical Communication*, *37*(1), 1-27. <https://doi.org/10.1177/10506519221121699>

Mollick, E. R., (July 11, 2016). Containing Multitudes: The Many Impacts of Kickstarter

Funding. [*SSRN*](file:///C:\Users\Shauna\Downloads\SSRN). <http://dx.doi.org/10.2139/ssrn.2808000>

Tian, J. (2021). Do You Want to Foresee Your Future? The Best Model Predicting the Success

of Kickstarter Campaigns. *ICMLC '21: Proceedings of the 2021 13th International Conference on Machine Learning and Computing*, 223-231. <https://doi.org/10.1145/3457682.3457716>

Tran, T., Dontham, M.R., Chung, J., & Lee, K. (2016). How to Succeed in Crowdfunding: A

Long-Term Study in Kickstarter. *ACM Transactions on Intelligent Systems and Technology, 0*(0), 0:0-0:28. <https://doi.org/10.48550/arXiv.1607.06839>