Deep dive into ANOVA(Analysis of Variance) Test

Van Nguyen
Department of Computer Science
Boise State University
vannguyen599@u.boisestate.edu

1 Introduction

Kickstarter is an online platform that enables creators to raise funds for their innovative projects from a community of supporters. The platform has become significantly popular among other crowdfunding platforms due to their user-friendly interface and dynamic ecosystem that fosters diverse creative campaigns. As someone who has backed several Kickstarter campaigns and found the experience satisfying, I at one point have considered launching some experimental project related to keyboards on the platform. Like other creators, the ultimate question is how to ensure their projects to be funded successfully. Given this question, we need to engage in a study to understand what factors can influence the success of a project. The objective for this course project is to analyze all Kickstarter project data to identify patterns in order to predict successful funding campaign. The motivation is to employ data science techniques to analyze and uncover actionable insights that can inform future crowdfunding strategies.

2 Methodology

The datasets used was generated by *https://webrobots.io/kickstarter-datasets/* which a platform where they have a regular scraper robot that crawls all Kickstarter projects every once a month. We used the datasets through 2025-10-13, which have 186,372 data points and 46 columns.

2.1 Data Exploration

Kickstarter does not provide any public API to retrieve data other then *https://status.kickstarter.com/api/v2/summary.json* which is used to look up the status of the platform. As a result, it is necessary to explore and investigate existing datasets from other platforms such as **Kaggle** to determine whether they are relevant or suitable to use. The selected dataset should be up-to-date and include all key factors required for an examination of funding trends and project success. we have attempted to scrape the data directly from the website but was not able to get all the sufficient data so we have decided to go with *https://webrobots.io/kickstarter-datasets/*. However to make sure that data is relevant and webrobots (Webrobots, 2025) does not just patch missing data with wrong information, we have matched data we scraped with what currently provided by webrobots on 5-6 projects. **sample.json** is the data that we have pulled down to check for webrobots's validity.

Other than that we need to go through the dataset in order to make sure that we only select relevant features to ensure effective data training. After consideration, we decided to go with the below list as features:

- The campaign's financial goal in USD
- The time taken to prepare the campaign (from the creation date to the launch date)
- The duration of the campaign (from the launch date to the deadline)
- The geographical location of the campaign (whether it is based in the US or not)
- The category under which the campaign falls

We need to do some data preprocessing where we dropped unused columns and check for duplicates or invalidate data points. Other than that for any timestamp we want to convert them to datetime object. In order to accommodate the above feature list, we need to create extra columns that suit our purpose namely *main_category*, *sub_category*, *campaign_goal_USD*, *campaign_location_US*, *campaign_timeline*, *prep_timeline*.

2.2 Data Analysis

We used the preprocess dataset to analyze the trends and evaluate how they affect the success

of a campaign. First we need to take a look at the success rate of all the campaigns in Figure 1. We can see that the success rate is roughly 60% while the failed rate is around 40% which is a not a significantly big gap. This makes us wondering what contributed to the gap between a successful and a failed project. To address this, we visually examine the influence of various features on project success.

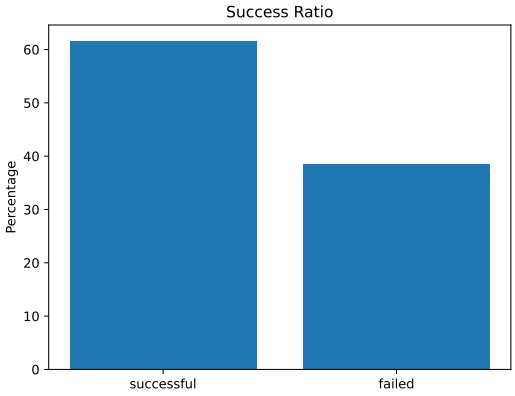

Figure 1: Success Ratio

We first look at the relationship between the campaign's financial goal and the success rate of it using box plot in Figure 2. Visually looking at the box plot graph, we can see that failed projects shows a massive number of outliers with very high goals up to $140,000,000. The primary takeaway from this plot is the significant difference in the magnitude and frequency of high-goal outliers between failed and successful projects. Projects with extremely large goals are overwhelmingly found in the failed group. This strong concentration of high-goal outliers in the failed state suggests that setting a very high financial goal might be strongly correlated with project failure.

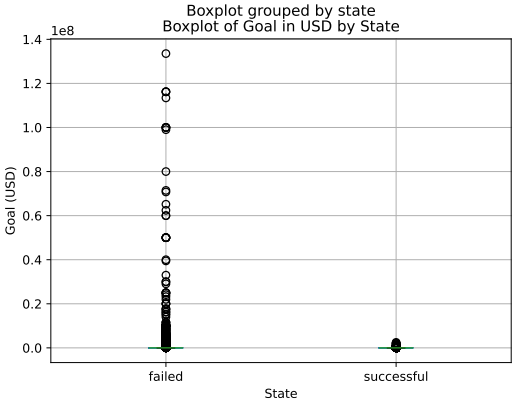

Figure 2: Goal in USD

Next, we take a look at the relationship between the campaign's geographical location and the success rate of it using Figure 3, and Figure 4. Based on the observations, the United States accounts for the majority of campaigns launched on Kickstarter, with over 60% of the total, and more than 66% of these campaigns are successful. The second-largest contributor to launched projects is Great Britain, accounting for approximately 10% of the total, with around 70% of these being successful. The disparity in the number of launched projects between the US and Great Britain is notably significant. The data suggests that launching a campaign in the United States is generally safer and more likely to succeed due to the higher volume and success rate of campaigns.

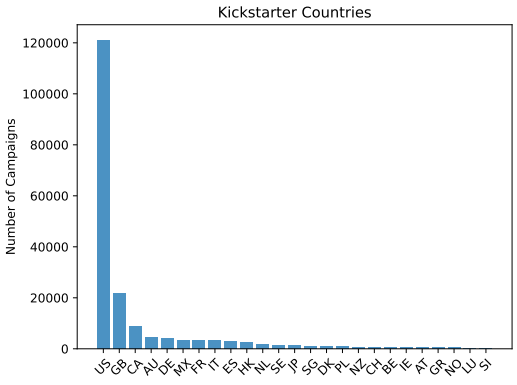

Figure 3: Campaign's country

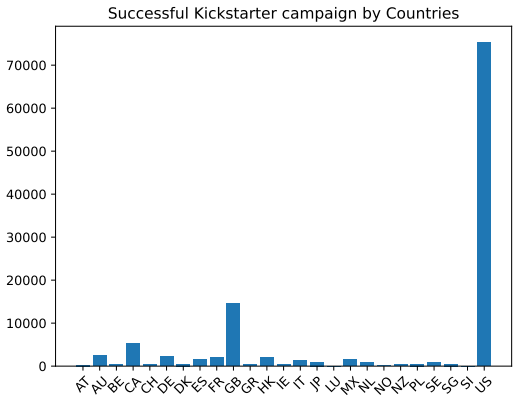

Figure 4: Successful rate by country

Next, we observe the relationship between the campaign's duration and the success rate of those projects in Figure 5. The majority of campaigns fall within the first three duration bins (0-60 days), with very few campaigns lasting longer than 60 days. The 20-40 day bin is overwhelmingly the most common campaign timeline, accounting for

over 100,000 projects. Despite having a moderate success rate, the 20-40 day bin also has the highest absolute count of failed campaigns ($\approx 70,000$) due to the sheer volume of projects launched in this window. The data suggests that a slightly longer 40-60 day timeline is associated with the highest likelihood of success

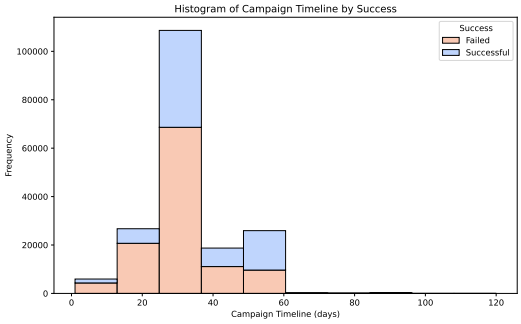

Figure 5: Histogram of Campaign Timeline by Success

Other than that, looking at the preparation in time in Figure 6 we can see that the distribution is heavily skewed, with the vast majority of all projects falling into the first bin. The shortest preparation timeline is by far the most popular choice for creators, reflecting a common behavior to launch soon after starting the project creation process.

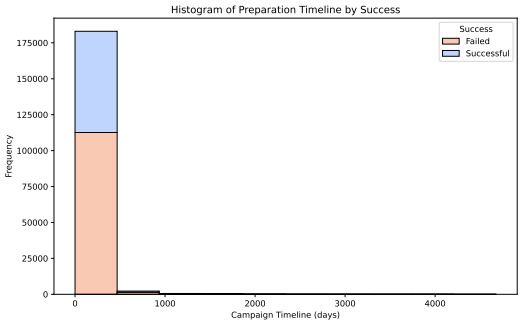

Figure 6: Histogram of Preparation Timeline by Success

Lastly, we analyze the relationship between category and the rate of success in Figure 7. The chart demonstrates a clear concentration of successful campaigns in the Arts, Media, and Entertainment sectors, with Music and Film & Video leading by a significant margin.

3 Experiments

In this section, we utilized the preprocessed data to train three different models: Logistic Regression, Decision Tree, and Random Forest, in order to determine the model that delivers the best performance. All three models are typically evaluated

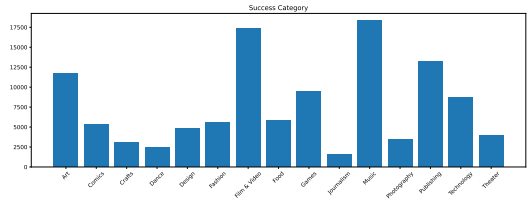

Figure 7: Success Category

using the same common classification performance metrics, which are designed to compare their predictions against the true labels. We initially trained the models using their default parameters to evaluate their performance, and then proceeded with hyperparameter tuning on the best-performing model. The table 1 summarized all the training metrics after model training process. The table showed that Logistic Regression and Random Forest had only slight differences across all evaluated metrics. However, Random Forest performed slightly better in both Accuracy and ROC.

Model type	accuracy	precision	recall	f1	roc
Logistic	0.690	0.701	0.866	0.774	0.712
Decision Tree	0.643	0.717	0.693	0.705	0.631
Random Forest	0.695	0.734	0.790	0.761	0.730

Table 1: Summary of chosen model's performance before tuning

In order to see it better, we also have a ROC curve (Developers, 2025) of all the three models in Figure 8. From the graph we can see that Random Forest model has the highest AUC at 0.734, indicating it is the overall best classifier among the three, as it has the best ability to discriminate between the positive and negative classes

3.1 Hyperparam tuning

We opted to fine-tune the Random Forest parameters using GridSearchCV due to its superior overall performance. Our goal was to reduce false positives and enhance precision. After param tuning, the best parameter that we got is criterion: entrophy and n_estimator: 500 which yields the following improvement in Table 2 We can see that accuracy has gone up 0.1 and precision has gone up 0.03 which is slightly better than what we have before tuning.

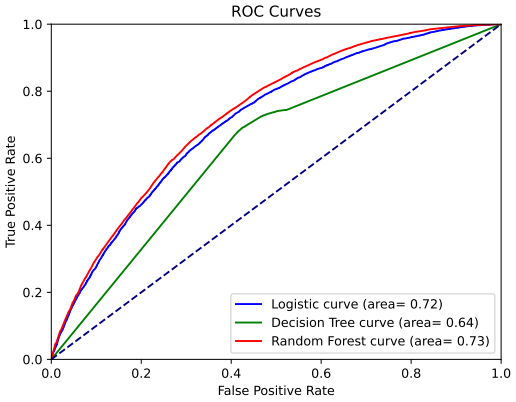

Figure 8: ROC Curves

accuracy	precision	recall	f1	roc
0.7007	0.7379	0.7968	0.76626	0.735

Table 2: Summary of chosen model's performance of Random Forest after tuning

3.2 Application

After selecting the model that we want, we exported the trained model to a file and then created simple web application with that model to predict whether or not a campaign will be successful or not. The website consisted a simple form that takes user inputs (features) and then evaluate the result with the Random Forest model. The application was fairly simple and was written using Flask that ran on port 5000 and have 2 routes with 1 POST Rest API.

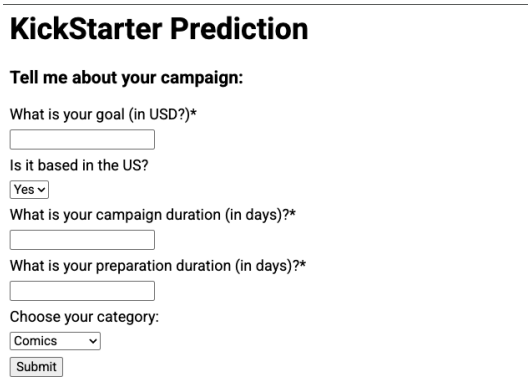

Figure 9: Web Application

4 Discussion

In this study, we analyzed Kickstarter campaign data to identify patterns and factors influencing the success of crowdfunding projects. By leveraging data preprocessing techniques, we extracted key features such as financial goals, campaign timelines, and geographical locations to better understand their impact on project outcomes. We have used three machine learning models—Logistic Regression, Decision Tree, and Random Forest—to evaluate their performance in predicting campaign success. Among these, Random Forest emerged as the best-performing model, demonstrating superior accuracy and ROC scores. Further hyperparameter tuning using GridSearchCV allowed us to optimize the Random Forest model, reducing false positives and improving precision. Our findings revealed several insights, including the correlation between high financial goals and campaign failure, the importance of campaign duration, and the dominance of certain categories and geographical regions in successful projects. These insights can guide future creators in designing effective crowdfunding strategies. Overall, this project highlights the value of data-driven approaches in understanding and improving crowdfunding outcomes. By applying machine learning techniques and statistical analysis, we demonstrated how predictive modeling can provide actionable insights for creators and stakeholders in the crowdfunding ecosystem.

References

Google Developers. 2025. Roc and auc - machine learning crash course. Accessed: 2025-12-01.

Webrobots. 2025. Kickstarter datasets. Accessed: 2025-12-01.