PREDICTING THE SUCCESS OF KICKSTARTER PROJECTS

Miss Shoon Lae Than Tun, Mr. Enes Talha Ciftci, Mr. Abdullakh Ali Unal

ICT Department of Rangsit International College (RIC)

**Corresponding author: shoonlae.t65@rsu.ac.th*

Abstract

Kickstarter has transformed how innovative ideas and projects receive funding by directly engaging a large audience through online platforms. Despite its popularity, only about 40% of projects on platforms such as Kickstarter succeed,leaving a significant portion of creators unable to bring their ideas to life within predetermined timeframes. This study investigates the factors that influence the success of Kickstarter projects and uses machine learning techniques to predict project outcomes.In this study, we used a dataset of Kickstarter projects, and the analysis shows how important early-stage campaign activities and temporal characteristics are in determining success. The study compares models such as Gradient Boosting and Logistic Regression, with 92% accuracy in predicting project outcomes. The findings provide valuable advice to project creators, demonstrating how early campaign activities, community engagement, and a well-planned strategy can significantly increase the likelihood of success. These insights assist creators in determining which factors are most important and offer practical steps for improving the effectiveness of crowdfunding campaigns.

Keywords*: Crowdfunding, Kickstarter, Project Success Prediction, Machine Learning, Campaign Management*

1. INTRODUCTION

Kickstarter has emerged as an effective tool for funding creative projects, democratizing the process of obtaining financial backing. Platforms such as Kickstarter allow individuals to connect directly with a large audience, removing traditional funding barriers. However, not all campaigns succeed, and identifying the factors that influence project success remains a significant challenge. Kickstarter projects must meet their funding targets to succeed, with outcomes classified as successful, canceled, or unsuccessful. The project category, funding goal, campaign duration, and quality of the project description all play important roles in determining success.

The purpose of this study is to analyze Kickstarter project data in order to identify patterns and key factors that contribute to project success. Machine learning is used to examine attributes such as project category, funding goal, campaign duration, and backer engagement. This study uses classification models to predict whether a project will succeed, providing creators with actionable insights for crafting more effective campaigns.

This study addresses the research question: *What factors contribute to the success of Kickstarter projects, and how accurately can machine learning models predict project outcomes?* Using a dataset of Kickstarter projects, this study applies Gradient Boosting and Logistic Regression to predict project outcomes, evaluating their effectiveness through metrics such as accuracy, precision, and recall. By focusing on structured project features, this research explores the potential of predictive analytics in the crowdfunding domain. The study evaluates model performance using metrics such as accuracy, precision, and recall to identify the most effective approach for predicting project success.

2. LITERATURE REVIEW

Crowdfunding has grown in popularity, with platforms such as Kickstarter allowing creators to bypass traditional funding methods. Previous research has identified project category, funding goal, and duration as important predictors of success. High-quality project descriptions and frequent updates have also been linked to improved results.

2.1 Machine Learning in Crowdfunding Prediction

Machine learning (ML) techniques have demonstrated their power in predicting outcomes across various domains, including crowdfunding. Supervised learning algorithms such as Naive Bayes and Random Forest are commonly used to determine whether a project is successful or unsuccessful. Naive Bayes is popular for classification tasks due to its simplicity and interpretability. Meanwhile, Random Forest is known for its robustness and ability to handle complex relationships between features like campaign category, funding goal, and duration.

Among advanced models, Logistic Regression is particularly valuable for crowdfunding prediction due to its balance of interpretability and performance. By modeling the probability of project success as a linear combination of features, it allows straightforward insights into the relationship between predictors like funding goals, campaign duration, and project categories. Logistic Regression is effective when relationships are linear, making it a popular baseline in classification problems.

Similarly, Gradient Boosting has emerged as a powerful model in crowdfunding prediction. By iteratively combining weak learners, Gradient Boosting captures intricate patterns in the data while minimizing prediction errors. Its ability to handle both categorical and numerical features makes it well-suited for crowdfunding datasets, where interactions between features like project descriptions and funding goals often determine outcomes. Despite its complexity, Gradient Boosting offers superior performance in scenarios where precision and recall are critical.

2.2 Factors Influencing Project Success

Several factors influence the success of crowdfunding campaigns, with project-specific characteristics and campaign strategies playing important roles. Studies have emphasized the importance of:

- **Project Category**: Certain categories, such as Technology and Games, perform better because of increased backer interest and community support.
- **Lower funding goals**: Lower funding goals associated with higher success rates because they are seen as more attainable.
- **Campaign Duration**: Campaigns should last between 30 and 45 days, since too short duration or too long duration can have a negative impact on backer interest.
- **Early Engagement**: Early activity, such as backer contributions and social media shares in the first few days, has been found to have a strong relation with overall success.

2.3 Comparison of Predictive Models

In crowdfunding success prediction, Logistic Regression and Gradient Boosting are widely used models. Logistic Regression is a simple, interpretable model that performs well when relationships between

features and success are linear. It provides clear insights into the probability of project success but can struggle with complex interactions between features.

Gradient Boosting, on the other hand, is a more advanced technique that builds strong models by combining weak learners. It excels at capturing non-linear patterns and interactions between features, resulting in higher accuracy. However, it tends to sacrifice interpretability for improved performance. Gradient Boosting is ideal for datasets where accuracy is critical, while Logistic Regression offers transparency and simplicity.

2.4 Research Gap

Existing studies on crowdfunding success prediction have provided valuable insights, but significant gaps persist. Most research emphasizes either structured data or unstructured data. Moreover, there is a scarcity of research focusing on models that balance predictive accuracy and interpretability, which are essential for providing actionable advice to project creators. This study aims to address these gaps by analyzing structured project features. By comparing models like Logistic Regression and Gradient Boosting , this research evaluates their effectiveness in delivering accurate predictions and interpretable results to support creators in designing successful campaigns.

3. Methodology

This study employed a systematic methodology to predict the success of Kickstarter projects. The process included data preprocessing, model development, and evaluation using Python Language. The methodology aimed to ensure accurate predictions and interpretable results, providing actionable insights for Kickstarter project creators.

3.1 Experimentation

The experimentation phase involved a structured approach to data preparation and modeling, which included:

1. **Dataset Acquisition**: The dataset was sourced from Kaggle's Kickstarter dataset, which included approximately 300,000 entries.
2. **Data Preprocessing**: The dataset was cleaned and filtered to process. Only rows with "Successful" and "Failed" values in the "State" column were retained, as the original dataset contained five categories. Remove unnecessary columns and make a new column called “duration”.
3. **Handling Missing Value**: Since there are some missing values in the dataset, we have to replace the missing value with frequent values and mean numbers for numeric values.
4. **Label Assignment**: The "State" column was designated as the target label for prediction tasks.
5. **Model Training**: Models were trained using Python Language, leveraging algorithms such as Logistic Regression and Gradient Boosting and Tensorflow for binary classification model and cross-validation.

3.2 Environment

The study was conducted using the following tools and environments:

- **TensorFlow**: For binary classification model training by cross-validation, and performance evaluation.
- **Python** : For data preprocessing, model training, and performance evaluation. Supplemented the process for any additional data exploration or visualization, leveraging libraries such as Pandas, NumPy, and Matplotlib.
- **Kaggle**: Source for the Kickstarter dataset used in this research.

3.3 Dataset Preparation

Data preprocessing was critical to ensure the dataset was clean, consistent, and suitable for modeling. The steps included:

- **Filtering**: Non-relevant states such as "Canceled," "Live," and "Suspended" were removed from the dataset.
- **Feature Engineering**: Created a new column, duration, by calculating the difference between the deadline and launched dates to represent the length of each project's campaign.
- **Feature Selection**: Key features such as project category, funding goal, and duration were retained for analysis.

3.4 Model Development

To predict the success of Kickstarter projects, model development was carried out by using classification models in Python. The following machine learning models were evaluated:

- **Logistic Regression**: A simple, interpretable model that is effective for linear relationships between features and project success.
- **Gradient Boosting**: A robust ensemble model that captures complex, non-linear patterns, offering high accuracy but lower interpretability.

3.5 Training and Testing

The dataset was divided into training and testing sets using a standard 70-30 ratio. Hyperparameter tuning was performed using GridSearchCV in Python to ensure optimal model performance. Each model's predictive capabilities were assessed using evaluation metrics such as accuracy, precision, recall, and F1 score. These metrics helped in comparing the performance of models Logistic Regression and Gradient Boosting, implemented using Python's scikit-learn libraries.

3.6 Model Evaluation

The models were evaluated based on their predictive performance across several metrics:

- **Accuracy**: The proportion of correctly classified instances.
- **Precision and Recall**: For both "Successful" and "Failed" categories, measuring the relevance of predictions and their completeness.
- **F1 Score**: A harmonic mean of precision and recall, ensuring a balanced evaluation.

This structured methodology allowed for a comprehensive analysis of model performance, contributing actionable insights for Kickstarter project creators.

4. RESULT AND DISCUSSION

The evaluation of two machine learning models—Gradient Boosting and Logistic Regression—showed that the Gradient Boosting had the highest accuracy (92%), followed by Logistic Regression (90%). Gradient Boosting achieved 95% recall for "Successful" but 88% for precision and for Logistic Regression, 82% for recall and 93% for precision. In contrast, the Gradient Boosting demonstrated a balanced performance in terms of precision and recall, making it the most reliable model for predicting project outcomes. Logistic Regression performed poorly in terms of accuracy and F1 score, indicating potential limitations in handling this dataset effectively.

The Gradient Boosting proved to be the most effective for predicting Kickstarter success due to its balance of precision and recall, providing creators with actionable insights. While Gradient Boosting had a high recall for "Successful," its uneven performance across classes limits practical utility. The relatively poor performance of Logistic Regression indicates the need for advanced tuning or feature engineering. Future research could improve these models by incorporating unstructured data, such as project descriptions, or by investigating ensemble techniques like Gradient Boosting, which can improve accuracy and robustness.

Empirical Result Table for Kickstarter Success Prediction

Model	Accuracy	Classification Error	Precision	Recall	F1 Score
Gradient Boosting	92.97%	7.03%	88%	95%	92%
Logistic Regression	90%	10%	93%	82%	87%

5. CONCLUSION

This study used a structured dataset of 300,000 entries to compare the performance of various machine learning models in predicting the success of Kickstarter projects. Among the models tested, the Gradient Boosting performed the best overall, balancing precision and recall to make reliable predictions. While Logistic Regression outperformed in recall for "Success," its uneven performance emphasizes the importance of choosing appropriate models for specific use cases. The findings highlight machine learning's potential to provide actionable insights for project creators, though additional improvements could be made by incorporating unstructured data and exploring advanced modeling techniques.

Acknowledgement

I would like to express my heartfelt gratitude to my professor and mentor for their invaluable guidance throughout this study. I also acknowledge the resources provided by Kickstarter, Kaggle and Google Colab, which were instrumental in completing this research.

References

- Andreoni, J. *Impure altruism and donations to public goods: a theory of warm-glow giving. The Economic Journal,* pages 464–477, 1990.
- Ashta, A., & Assadi, D. *Do social causes and social technology meet? Impact of web 2.0 technologies on peer-to-peer lending transactions. Cahiers du CEREN,* 29:177–192, 2009.
- Bennett, S. *Log-logistic regression models for survival data. Applied Statistics,* pages 165–171, 1983.
- Bruett, T. *Cows, kiva, and prosper.com: How disintermediation and the internet are changing microfinance. Community Development Investment Review,* 3(2):44–50, 2007.
- Chapelle, O. *Modeling delayed feedback in display advertising.* In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pages 1097–1105, ACM, 2014.
- Choo, J., Lee, D., Dilkina, B., Zha, H., & Park, H. *To gather together for a better world: Understanding and leveraging communities in micro-lending recommendations.* In *Proceedings of the 23rd International Conference on the World Wide Web,* pages 249–260, 2014.
- Cox, D. R. *Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological),* pages 187–220, 1972.
- Etter, V., Grossglauser, M., & Thiran, P. *Launch hard or go home!: Predicting the success of Kickstarter campaigns.* In *Proceedings of the First ACM Conference on Online Social Networks,* pages 177–182, 2013.
- Galak, J., Small, D., & Stephen, A. T. *Microfinance decision making: A field study of prosocial lending. Journal of Marketing Research,* 48(SPL) –S137, 2011.
- Gerber, E. M., Hui, J. S., & Kuo, P.-Y. *Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms.* In *CSCW Workshop,* 2012.