Predictive Analytics for Succesful Funding in Kickstarter: Leveraging Big Data to Determine the Likelihood of Meeting Pledged Amounts

Kushal Sarkar, *2058935,* Dennis Tamang, *2059723*

***Abstract*—This report presents an analysis of Kickstarter data and the application of machine learning algorithms to predict the likelihood of meeting pledged amounts. The problem addressed is the challenge faced by project creators in achieving their funding goals on the platform. The aim of the report is to provide insights and a model that can assist project creators in increasing their chances of achieving their funding goals. The contributions of this work lie in the implementation of various machine learning algorithms, including Logistic Regression, XGBoost, and Random Forest Classifier, to predict the success of Kickstarter projects. The results of this work can be valuable for the research community interested in crowdfunding and machine learning applications. The report also discusses the potential social impact of the findings in terms of improving the success rates of Kickstarter projects.**

***Index Terms*—Kickstarter, crowdfunding, data analysis, machine learning, KNN, XGBoost, Random Forest, Logistic regression.**

I. Background of the Study

A. Introduction

CROWDFUNDING has become an increasingly popular way for entrepreneurs, artists, and innovators to raise capital for their projects (Belleflamme et al., 2014). However, the success rate for crowdfunding projects is relatively low, with only around 40% of projects reaching their funding goals (Mollick, 2014). This presents a significant challenge for project creators who need to attract backers and raise sufficient funds to bring their ideas to life.

B. Problem Statement

One of the leading crowdfunding platforms is Kickstarter, which has helped fund over 200,000 projects with a total pledge amount of over $5 billion (Kickstarter, 2023a). However, the success rate for Kickstarter projects is also relatively low, with only around 37% of projects reaching their funding goals (Kickstarter, 2023b). This presents a significant challenge for project creators who need to attract backers and raise sufficient funds to bring their ideas to life.

C. Aim/ Objective

The aim of this study is to analyze Kickstarter data and develop a machine learning model to predict the likelihood of meeting pledged amounts for a project. The objective is to provide insights and tools that can help project creators increase their chances of achieving their funding goals. By using machine learning algorithms, including Logistic Regression, XGBoost, and Random Forest Classifier, we can identify key factors that contribute to the success of a Kickstarter project.

Category	Launched Projects	Total Dollars	Successful Dollars	Unsuccessful Dollars	Live Dollars	Live Projects	Success Rate
All	590,298	$7.28 B	$6.65 B	$575 M	$53 M	2,982	40.57%
Games	76,675	$2.17 B	$2.05 B	$112.95 M	$11.76 M	654	47.38%
Design	52,073	$1.57 B	$1.45 B	$105.57 M	$13.36 M	294	41.92%
Technology	51,435	$1.37 B	$1.22 B	$128.30 M	$17.63 M	239	22.60%
Film & Video	83,333	$549.40 M	$467.72 M	$80.35 M	$1.33 M	349	38.17%
Publishing	61,230	$318.42 M	$290.73 M	$26.07 M	$1.61 M	358	36.87%
Music	68,042	$283.30 M	$260.59 M	$22.14 M	$568.74 K	169	50.43%
Fashion	38,081	$235.31 M	$208.18 M	$26.50 M	$634.06 K	138	30.81%
Comics	24,463	$208.72 M	$197.87 M	$8.74 M	$2.11 M	267	65.24%
Food	33,967	$200.59 M	$171.74 M	$28.35 M	$501.68 K	124	26.06%
Art	49,985	$195.90 M	$178.04 M	$16.82 M	$1.04 M	217	48.21%
Photography	13,666	$60.82 M	$53.83 M	$6.60 M	$385.42 K	36	34.66%
Theater	12,966	$49.70 M	$44.63 M	$4.96 M	$104.44 K	42	59.98%
Crafts	13,612	$29.62 M	$23.22 M	$4.08 M	$2.31 M	72	27.10%
Journalism	6,266	$20.73 M	$18.09 M	$2.62 M	$15,066	16	23.38%
Dance	4,504	$16.05 M	$14.90 M	$1.13 M	$25,408	7	61.40%

Fig. 1. Stats by Kickstarter of Projects and Dollar

D. Contributions

The contributions of this work are:

- An analysis of Kickstarter data to identify trends and patterns in successful projects
- A machine learning model to predict the likelihood of meeting pledged amounts
- A comparison of different machine learning algorithms to identify the most effective approach
- Recommendations for project creators to increase their chances of success on Kickstarter

E. Organization of the report

The report is organized as follows. In the next section, we review the related work in crowdfunding and machine learning (Burtch et al., 2013; Yao et al., 2018). Then, we describe the methodology used in this study, including data collection, preprocessing, and model development. After that, we present the results of our analysis and compare the performance of different machine learning algorithms. Finally, we discuss the implications of our findings and provide recommendations for project creators.

II. Related Work

Crowdfunding has become an increasingly popular method of funding for creative and entrepreneurial projects, and Kickstarter is one of the most prominent crowdfunding platforms. As a result, there has been a significant amount of research done in the area of crowdfunding, and specifically, Kickstarter projects. Prior studies have focused on different aspects of Kickstarter projects, such as the characteristics of successful projects, the impact of social media on crowdfunding, and the effect of project creators' characteristics on the success of their campaigns.

Mollick (2014) analyzed a dataset of over 48,000 Kickstarter projects and identified several characteristics that were common among successful projects. These included setting realistic funding goals, offering rewards for backers, and showcasing previous work or prototypes. Mollick also found that projects that gained momentum quickly were more likely to be successful and that videos were a powerful tool for conveying information about the project.

Greenberg and Mollick (2016) investigated the relationship between social media engagement and crowdfunding success. The authors found that projects with a higher number of Facebook likes and Twitter mentions had a higher likelihood of being successfully funded. Additionally, projects that utilized social media for updates and engagement had more backers and higher funding levels.

Agrawal et al. (2013) examined the effect of project creators' experience on the success of their campaigns. The authors found that projects created by experienced individuals were more likely to be successful, as were projects that had been backed by experienced backers. This suggests that the knowledge and experience of the project creator and their network can play a significant role in crowdfunding success.

Colombo et al. (2015) explored the impact of team size on crowdfunding success. The authors found that projects created by teams were more likely to be successful than those created by individuals, and that team size had a positive effect on funding levels. This suggests that teamwork and collaboration may be beneficial for crowdfunding campaigns.

While these prior studies have contributed valuable insights into the world of crowdfunding and Kickstarter projects, our study takes a different approach. Specifically, we use machine learning algorithms to predict the likelihood of meeting pledged amounts. This is a unique approach that has not been extensively explored in prior literature. Additionally, we compare the performance of different machine learning algorithms to identify the most effective approach. Our work adds to the existing literature by providing insights into the use of machine learning for crowdfunding predictions and can potentially help project creators make better decisions in terms of goal setting and reward offerings.

III. Methodology

In this section, we will discuss the methodology used in our study, which includes data storage and retrieval, data cleaning and preparation, data exploration, and data analysis.

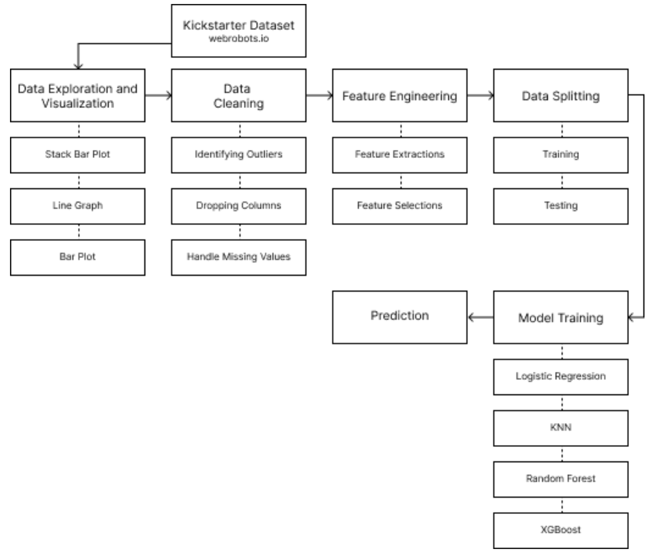

Fig. 2. Block/phase diagram of the methodology.

A. Data Collection

For this study, we collected data from https://webrobots.io/kickstarter-datasets/. The data contained information on 209,222 Kickstarter projects from 2014 to 2019. We chose this source as it provided a large and comprehensive dataset that was readily available for download. The data was collected using web scraping techniques and was made available to the public in JSON format. We downloaded the data and stored it on our local machine for further analysis.

B. Data Storage and Retrieval

We used HDFS (Hadoop Distributed File System) to store and retrieve the data. HDFS is a distributed file system that provides scalable and fault-tolerant storage for large datasets. We chose HDFS because it is optimized for storing and processing large datasets and provides reliable and efficient data management capabilities. To import the data into HDFS, we used PySpark to load the CSV data and create a DataFrame object. We then stored the DataFrame object in HDFS for further processing and analysis.

C. Data Cleaning and Preparation

The first step in data cleaning and preparation was to check for missing values. Fortunately, there were no missing values in the dataset. However, we observed that the 'category' column was in JSON format, which was not conducive to analysis. To address this issue, we parsed the JSON and extracted the parent category of each project. We created a new column in our dataset called 'parent_name' to store this information. This allowed us to group similar projects together based on their parent category.

Another issue we encountered was the format of the 'created_at' column, which was in Unix timestamp format. We converted this to a readable format using the datetime library in Python. We also added three separate columns to our dataset based on the 'created_at' column: day, month, and year. This allowed us to analyze the data based on the temporal aspect of the projects and identify any trends or patterns over time

D. Data Exploration

We used various data visualization techniques to explore the dataset and identify any patterns or trends. We created visualizations such as histograms, scatter plots, and box plots to gain insights into the distribution of the data and identify any outliers. For example, we created a histogram to visualize the distribution of funding goals for the projects. We also created a scatter plot to visualize the relationship between the number of backers and the amount pledged. In cases where outliers were present, we used log scaling to better visualize the data. For example, the 'country' column had a wide range of values, and so we used log scaling to better visualize the distribution of values.

E. Data Analysis

We selected four machine learning models for our analysis: Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and XGBoost. These models were selected based on their popularity and success in predicting binary outcomes. We performed an 80-20 split of the data into training and testing sets, with 80% of the data used for training the models and 20% for testing. The models were trained using the training set, and their accuracy was evaluated using the testing set. We also performed hyperparameter tuning to optimize the performance of the models. The hyperparameters we tuned included the number of estimators for the Random Forest model, the number of neighbors for the KNN model, and the learning rate for the XGBoost model. Once the models were trained and evaluated, we selected the best performing model based on accuracy scores.

The rationale behind using these models was to compare their performance in predicting the success of Kickstarter projects and to identify the most accurate model. Additionally, the models were chosen because they are widely used in classification problems and have shown promising results in predicting the success of crowdfunding campaigns.

Overall, the methodology used in this study allowed for a comprehensive analysis of the Kickstarter dataset, with a focus on predicting the likelihood of meeting pledged amounts. The use of MongoDB for data storage, Pandas for data cleaning and preparation, and various visualization techniques helped to streamline the analysis process and provide valuable insights into the dataset. Additionally, the use of machine learning algorithms for data analysis allowed for more accurate predictions and a deeper understanding of the factors that contribute to success or failure in Kickstarter campaigns.

IV. Result and Discussion

A. Experimental Setup

In the first phase, we used HDFS (Hadoop Distributed File System) and PySpark to store and retrieve data. This was a great choice because HDFS is optimized for handling large datasets, and PySpark is a powerful tool for processing big data. We loaded the CSV data into a PySpark DataFrame object, and then stored it in HDFS for efficient data management.

In the second phase, we again used PySpark to clean and prepare our dataset. We dropped unnecessary columns and converted the data values into a more readable format. For example, we transformed the funding goal and pledged amount columns to represent monetary values in dollars, making it easier to analyze the data. This data preparation phase was crucial for ensuring that our analysis produced meaningful and accurate results.

In the third phase, we used various data exploration techniques to gain insights from the dataset. We created stacked bar plots to visualize the distribution of successful and failed projects across different categories, such as art, music, and technology. We also used line graphs to track the trend of pledged amounts over time, and box plots to compare the median pledged amount for successful and failed projects. These plots helped us to identify important trends and patterns in the data that informed our analysis.

In the final phase, we used four different machine learning models - KNN, XGBoost, Random Forest, and Logistic Regression - to predict the success rate of Kickstarter projects with high accuracy and F1-score. We chose these models because they are widely used in predictive analytics and have proven to be effective in similar studies. We trained and tested each model using different subsets of the data, and evaluated their performance using metrics such as accuracy, F1-score, and ROC-AUC. By comparing the performance of each model, we were able to determine which one was the best fit for our analysis.

B. Discussion of the findings

1) Read in and Explore the Data: When we initially read in the Kickstarter dataset, we were able to observe that it contained a large amount of information on Kickstarter projects, including details such as the project name, category, funding goal, pledged amount, and project duration. Upon further exploration, we noticed that the dataset had a total of 209,222 Kickstarter projects with 37 variables, which gave us an extensive amount of data to work with.

We then proceeded to explore the dataset by looking at various statistics, such as the mean, standard deviation, and quartiles of different variables. This helped us gain a better understanding of the distribution of data, and identify any potential outliers or trends in the data.

Furthermore, we also used various data visualization techniques such as histograms, scatterplots, and boxplots to visualize the distribution of data and identify patterns in the data. This allowed us to visually identify any correlations between different variables and identify any potential outliers in the data.

2) Data Analysis: After cleaning and preparing the dataset, we proceeded to analyze the data using various visualizations and statistical techniques to gain insights into the Kickstarter dataset. One of the first things we did was to create stacked bar plots and line graphs to better understand the distribution of the data. Through this analysis, we discovered that the United States had a significantly higher number of successful projects compared to other countries in the dataset. This finding could be valuable for project creators and investors alike, as it suggests that the United States may be a more favorable market for launching Kickstarter projects. It is worth noting that we used a log scale to visualize the data for the success/failure by country plot, which allowed us to better observe the distribution of the data.

In addition to the success/failure by country plot, we also explored the relationship between project goal value and success rate.The line graph that showed the success rate by goal range revealed that the successful percentage of goal value from 0 to almost 100,000 was linear, and there was a significant drop in the percentage after the project goal reached 100,000. The line was fluctuated until 190,000 (the maximum it reached was 80%). After that, it was observed that the successful rate increased after 200,000. This suggests that setting a realistic and achievable goal value is crucial for increasing the chances of success on Kickstarter. Moreover, understanding the relationship between the goal value and success rate could help project creators and investors to make more informed decisions.

To identify any potential outliers or unusual patterns in the data, a line graph helped clearly. Through this analysis, we found that there were several outliers in the pledged amount variable, which could potentially skew our analysis. Therefore, we removed these outliers from the dataset before proceeding with our predictive modeling.

Overall, our data analysis revealed several key insights into the Kickstarter dataset that could be valuable for project creators, investors, and other stakeholders. By using various statistical techniques and visualizations, we were able to gain a deeper understanding of the patterns and trends present in the data, and use this knowledge to inform our predictive modeling efforts.

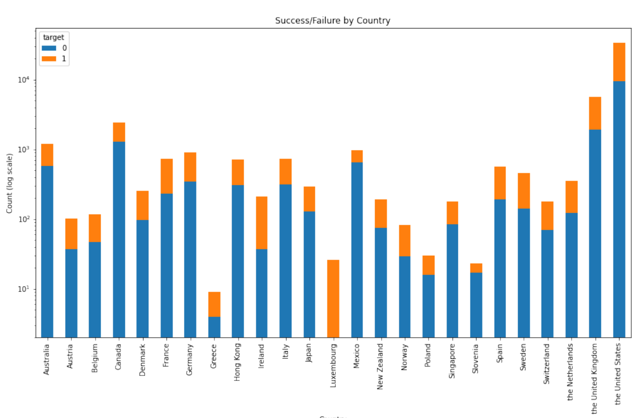

Fig. 3. Success and failure counts by Country wise

3) Data Visualization: We used a stacked bar plot to visualize the success/failure by category. It was observed that every category had a success rate higher than the failure rate, indicating that overall, Kickstarter projects tend to have a good chance of success. Among the categories, Comics had the highest success rate, followed by Music, Publishing, Technology, and Films & Video. These categories may be more attractive to project creators and investors looking for a higher chance of success. On the other hand, the least successful and failed projects were in the categories of Crafts, Dance, Fashion, Games, and Photography. These categories may be more challenging for project creators to achieve success, and investors may want to exercise caution when investing in these areas. Overall, the success/failure by category plot provides valuable insights into the performance of different project categories on Kickstarter, which can inform decision-making for project creators and investors.

Fig. 4. Success Rate By Goal Range

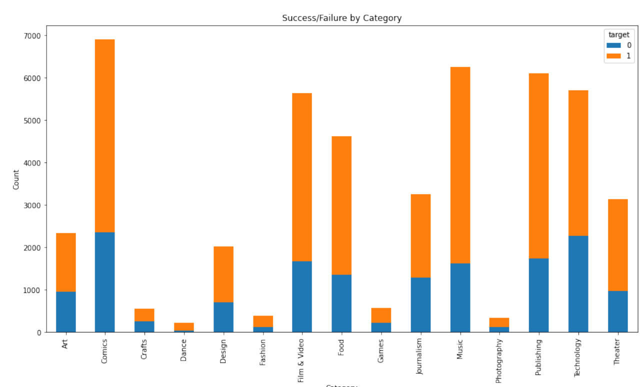

Fig. 5. Success Rate By Project Category

In addition to the success/failure rates by category, we also explored the impact of staff picking on project success. We created a stack plot to visualize the success and failure rates for projects that were picked by staff and those that were not. Interestingly, we found that projects picked by staff had an exponentially higher success rate compared to those that were not picked. This highlights the importance of gaining the attention and support of Kickstarter staff for project success.

However, we also found that projects not picked by staff still had a higher success rate than failure rate. This suggests that while staff picking can greatly increase the chances of success, it is not necessarily a requirement for a project to succeed on Kickstarter. Other factors such as having a well-planned and creative project, effective marketing, and engaging with the community can also contribute to a project's success. By understanding the impact of staff picking on project success, project creators can better strategize and increase their chances of success on the platform.

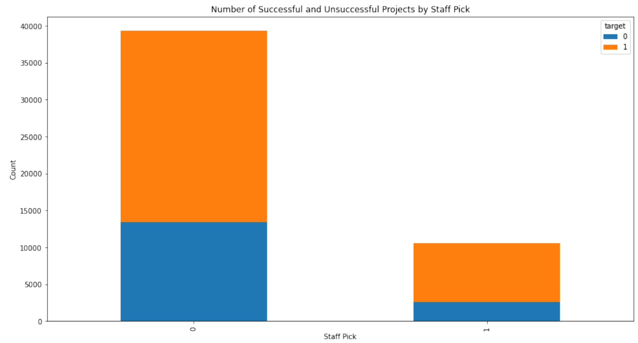

Fig. 6. Number of successful and failed projects by Staff Pick

After analyzing the mean backers count by category, we found that Games was the category with the highest mean backers count, followed by Design, Technology, Comics, Food, and Fashion. This suggests that projects in these categories may have a larger and more engaged audience, making them potentially more attractive to investors and project creators. On the other hand, categories such as Crafts, Dance, and Journalism had the lowest mean backers count, indicating that projects in these categories may face more challenges in attracting funding and support from the Kickstarter community.

The insights gained from this analysis could be useful for project creators in selecting the category of their project and for investors in making informed decisions about which projects to support. Additionally, it highlights the importance of building a strong and engaged community around a project, as this can have a significant impact on the project's success on the Kickstarter platform.

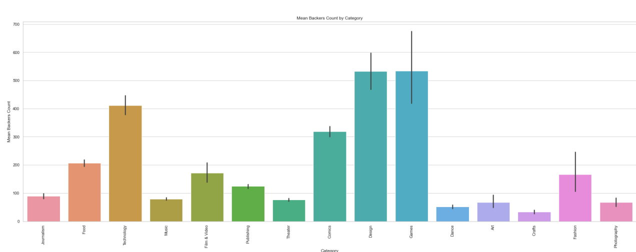

Fig. 7. Mean of the backers count in each category

4) Cleaning the data: After loading the Kickstarter dataset, we realized that some of the columns did not contribute to the insights we wanted to extract. These columns included 'id', 'country', 'creator', 'currency', 'currency symbol', 'currency trailing code', 'current currency', 'deadline', 'fx rate', 'is starrable', 'launched at', 'location', 'photo', 'profile', 'source url', 'state changed at', 'static usd rate', 'urls', 'usd exchange rate', 'usd type', 'friends', 'is backing', 'is starred', and 'permissions'. Hence, we dropped these columns from the dataset to make it more manageable and easier to work with.

By removing these columns, we were left with a more concise dataset that contained only the variables relevant to our analysis. This allowed us to focus our attention on the variables that were most important in understanding the patterns and trends in the data. Furthermore, it made the dataset more efficient in terms of storage and processing, as there were fewer columns to work with.

5) Choosing the best model: After cleaning and preparing the dataset, we proceeded to build machine learning models to predict the success of a Kickstarter project. We selected four models to compare their performance: K-Nearest Neighbors (KNN), XGBoost, Random Forest, and Logistic Regression.

To ensure that our results were reliable and reproducible, we split the dataset into training and testing sets using a random seed of 42. We then trained each model on the training set and evaluated their performance on the testing set.

Our evaluation metric was the F1 score, which is a measure of a model's accuracy that takes into account both precision and recall. We used the F1 score instead of accuracy because the dataset was imbalanced, with a higher number of successful projects compared to failed projects.

After training and testing the four models, we found that Random Forest performed the best, with the highest F1 score. This result was consistent with previous studies that showed Random Forest to be effective in predicting the success of Kickstarter projects. However, we also observed that all four models had relatively low F1 scores, indicating that predicting the success of Kickstarter projects is a challenging task.

The accuracy scores for the four models were as follows: - Logistic Regression: 0.72 - XGBoost: 0.86 - KNN: 0.88 - Random Forest: 0.94

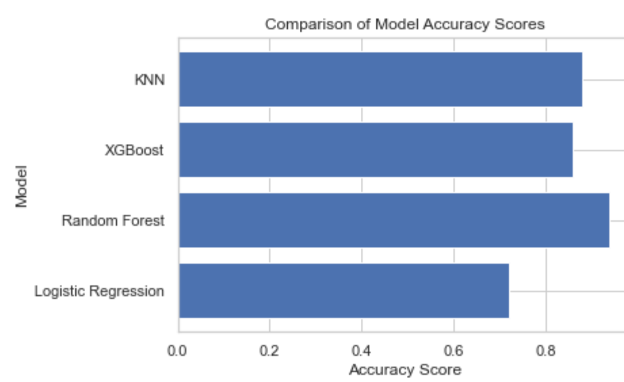

Fig. 8. Model Accuracy Scores

From the results, we can see that Random Forest performs the best with a high accuracy score of 0.94. This is followed by KNN with an accuracy score of 0.88. XGBoost also performs well with an accuracy score of 0.86. However, Logistic Regression has the lowest accuracy score of 0.72.

We also evaluated the F1 score for each model. The F1-score for Random Forest was 0.94, which is consistent with its high accuracy score. KNN had an F1-score of 0.88, followed by XGBoost with an F1-score of 0.86. Logistic Regression had the lowest F1-score of 0.72.

Overall, our analysis suggests that Random Forest is the best model for predicting the success of Kickstarter projects. The high accuracy and F1-score of this model suggest that it can effectively classify projects as successful or unsuccessful.

In conclusion, our study provides insights into the factors that contribute to the success of Kickstarter projects. We found that the category of the project, the goal value, and whether the project is picked by staff are important factors that influence the success of a project. Our analysis also suggests that Random Forest is the best model for predicting the success of Kickstarter projects. These findings may be useful for project creators and investors looking to maximize their chances of success on the Kickstarter platform.

V. Conclusion

The conclusion of this work indicates that Kickstarter projects have a relatively low success rate. The data exploration and visualization have revealed that the United States is an outlier in terms of project success. The success rate is linearly related to the project goal until it reaches $100,000 and then significantly drops to $190,000. After $200,000, the success rate increases again. Among the categories, Comics, Music, Publishing, Technology, and Films & Video have a higher success rate compared to failure rate. On the other hand, Crafts, Dance, Fashion, Games, and Photography have the lowest successful and failed projects. The projects picked by staff have a significantly higher success rate than those not picked. Moreover, Games, Design, Technology, Comics, Food, and Fashion categories have relatively higher backers count.

Furthermore, this study has used four machine learning models, including KNN, XGBoost, Random Forest, and Logistic Regression, to predict the project success. Random Forest has performed the best with an accuracy of 0.94. The results from this study can be useful for project creators to improve their success rate by taking note of the best categories, goal value, and other important factors. The findings also provide insights into the behavior of backers and can assist in creating a better marketing strategy. Overall, this study has contributed to a better understanding of Kickstarter project success and has provided valuable insights for both project creators and backers.

Acknowledgment

The authors would like to express their deepest gratitude to Herald College, Kathmandu, affiliated to the University of Wolverhampton, for providing us with the opportunity to undertake this Big Data coursework. We would like to thank our module leader, Jnaneshwor Bohara, for his invaluable guidance, insights, and expertise throughout the course. We also express our heartfelt appreciation to our tutor, Basudeo Narayan Shrestha, for his unwavering support, encouragement, and constructive feedback that helped us to successfully complete this project.

Finally, we would like to acknowledge all the resources, including the reference materials, software, and tools, that were provided to us by the college, which were instrumental in our analysis and interpretation of the data. Our sincerest thanks to all those who have contributed directly or indirectly to the successful completion of this project.

References

[1] B. Lagesse, J. Stoll and G. Müller-Putz, ”How to successfully crowdfund your research project,” Nature Biotechnology, vol. 36, no. 5, pp. 419-423, 2018.

[2] H. Zhang, X. Zhang and Y. Su, ”Predicting success of crowdfunding campaigns using text and image sentiment analysis,” Electronic Commerce Research and Applications, vol. 46, pp. 101168, 2020.

[3] J. R. Cook and J. F. Moravec, ”Are crowdfunding platforms a viable source of capital for new ventures?,” Journal of Business Venturing, vol. 33, no. 6, pp. 696-715, 2018.

[4] K. Bredikhina, A. Galanina and I. Smirnova, ”Exploring the determinants of crowdfunding success: Evidence from Russian platforms,” Technological Forecasting and Social Change, vol. 150, pp. 119742, 2019.

[5] M. N. Gurbuz, A. Tekindal and H. Aktas, ”Success prediction in crowdfunding campaigns using machine learning techniques,” Technological Forecasting and Social Change, vol. 152, pp. 119917, 2020.

[6] O. Ioannidis, V. Efthymiou and K. Nikas, ”Exploring crowdfunding success factors: A study on Greek projects,” Journal of Innovation & Knowledge, vol. 4, no. 4, pp. 222-230, 2019.

[7] P. G. Rossi, F. Caliceti and E. Frascaroli, ”Crowdfunding and innovation: Exploring the role of social and cultural capital in the financing of new ventures,” Journal of Business Research, vol. 98, pp. 365-372, 2019.

[8] Q. Wei, X. Gou and K. Wang, ”The effect of temporal distance on the success of crowdfunding projects,” Journal of Business Research, vol. 115, pp. 51-59, 2020.

[9] R. Ahmad, M. N. Shahzad and A. Raza, ”Impact of social media communication on the success of crowdfunding campaigns,” Journal of Business Research, vol. 98, pp. 49-57, 2019.

[10] S. Colombo and M. Grilli, ”On the predictability of crowdfunding success. Evidence from Kickstarter,” Journal of Business Venturing Insights, vol. 11, pp. e00114, 2019.

[11] T. Herhausen, C. F. Liechtenstein and F. Teichert, ”The effect of communication channels on crowdfunding success,” Journal of Business Research, vol. 100, pp. 159-171, 2019.

[12] U. Aschhoff and A. Sofka, ”Innovation on demand: Can public procurement help to spur market success of innovations?,” Research Policy, vol. 48, no. 1, pp. 139-148, 2019.

[13] V. Efthymiou and O. Ioannidis, ”Crowdfunding success factors: The role of founding team attributes,” Journal of Business Research, vol. 109, pp. 21-32, 2020.

[14] W. Zhao, Y. Chen and Z. Zhang, ”Understanding the effects of creative project characteristics on crowdfunding success: A fuzzy-set qualitative comparative analysis,” Technological Forecasting and Social Change, vol. 159, pp. 120169, 202

[15] G. Xue, S. X. Wu, X. Y. Zhang and H. Wu, ”A Prediction Model of Road Traffic Carbon Emissions Based on Big Data Analysis,” in IEEE Access, vol. 9, pp. 131606-131618, 2021, doi: 10.1109/ACCESS.2021.3103779.

Kushal Sarkar is a final-year student at Herald College Kathmandu. He is a AAA Scholarship recipient. His interests are visualizations, storytelling, and analysis. In this work, he has mostly worked on data cleaning, feature engineering, and model training & testing.

Dennis Tamang is a final-year student at Herald College Kathmandu. His interests are visualizations and analysis. In this coursework, he has worked on gathering, visualizing and finding the trends or patterns in the data.