Data Preprocessing

The study initiated with data preprocessing on the Kickstarter dataset. The focus was on projects classified as 'successful' or 'failed,' ensuring relevance to the study's objectives. The first step involved handling missing values, a critical task to maintain data integrity. For instance, projects missing essential information like 'name' were excluded, recognizing the importance of complete data for accurate analysis. Categories with missing entries were filled with 'Unknown,' maintaining categorical continuity. A crucial aspect of preprocessing was the removal of projects with illogical financial goals (negative or zero funding goals). Feature engineering played a pivotal role in enhancing the dataset's potential. Time-related features were extracted, such as the month, day, and year of project creation and launch. Furthermore, the duration between the project's creation and launch was calculated, offering insights into the project preparation phase. Also, Binary Encoding was applied to the arget variable {'failed': 0, 'successful': 1}and one-hot encoding was applied to variables like 'country', 'currency', and 'category' for future modeling.

Model Development, Evaluation, and Results

Classification Model: The classification model was developed using RandomForest and GradientBoosting algorithms. The model's design was mindful of using predictors available at the project's launch, thereby ensuring its real-world applicability. The model underwent thorough evaluation based on accuracy, precision, recall, and F1-score metrics, providing a comprehensive understanding of its performance. The RandomForest model demonstrated respectable accuracy (0.76) and cross-validation scores, indicating its reliability. However, the GradientBoosting model slightly outperformed RandomForest, showcasing higher accuracy

(0.77) and stability across different datasets. This superiority suggests the GradientBoosting model's enhanced ability to generalize and its potential effectiveness in practical scenarios.

Clustering Model: The K-Means clustering algorithm was applied to the standardized features, resulting in the formation of three distinct clusters. Each cluster represents a unique grouping of Kickstarter projects:

1. Cluster 0: This cluster seems to represent large-scale, ambitious projects, likely necessitating elaborate planning, substantial resources, and extensive marketing strategies. The extended preparation period might involve detailed project development, comprehensive market research, and building a strong promotional campaign.
2. Cluster 1: This cluster includes projects with lower average goals (around $14,207) and pledged amounts (approximately $32,543). The projects here are characterized by shorter durations between creation to launch (averaging about 29 days) and from launch to deadline (approximately 28 days).
3. Cluster 2: Projects in this cluster set significantly higher goals (averaging around $275,976) and achieve substantial pledged amounts (average close to $1,119,662). They also exhibit longer periods from creation to launch (about 68 days) and moderately long campaign durations (approximately 38 days).

Business Implications and Practical Applications

The GradientBoosting Classification Model, achieving an accuracy of 77.37%, offers substantial insights for the Kickstarter community. For project creators, it serves as a crucial decision-making tool, providing early predictions of a project's success likelihood. This can guide them in refining their project concepts or adjusting their strategies. Investors and backers can utilize the model to identify promising projects, reducing the risk associated with funding potentially unsuccessful campaigns. Furthermore, Kickstarter can integrate this model into their platform to enhance project visibility and success rates, aligning with user interests and market trends. For market researchers, the model's feature importance analysis can reveal key factors that drive crowdfunding success, offering valuable insights for future project development and marketing strategies.

The clustering model's insights reveal the diverse strategies and dynamics of Kickstarter projects, underscoring the varied approaches needed for different types of crowdfunding campaigns. Cluster 0, with its large-scale, ambitious projects, suggests a market segment where extensive planning, resource allocation, and sophisticated marketing are key to success. Such projects may appeal to a broad audience and require a sustained effort in engagement and promotion. On the other hand, Cluster 1, featuring projects with modest goals and shorter campaign durations, highlights a segment where agility, niche targeting, and community engagement can be pivotal. These projects might appeal more to specific interest groups or rely on strong community support. Cluster 2, characterized by high goals and significant funding, indicates a unique subset of projects that, despite their ambitious targets, manage to secure substantial backing, possibly due to their innovative nature or strong market demand. This segmentation offers valuable guidance for project creators in tailoring their strategies according to their project type and for backers in identifying projects that align with their investment preferences or interests. For the Kickstarter platform, these insights can inform the development of targeted support and resources, enhancing the success potential of diverse project types.