Title: Comprehensive Analysis of Kickstarter Project Success Using Machine Learning

I- Introduction:

This report offers a focused analysis of the Kickstarter dataset, employing machine learning techniques to predict project outcomes and discern patterns. Our approach encompasses thorough data preparation, the deployment of various classification models, and clustering algorithm exploration. The classification segment aims to predict project success, crucial for creators in strategizing their campaigns. Meanwhile, the clustering analysis seeks to uncover project groupings, providing insights for platform strategists. Overall, this study aims to extract actionable insights, benefiting both project creators and platform managers in the dynamic crowdfunding environment.

II- Data Preparation and Feature Engineering

In the initial phase, the dataset underwent rigorous cleaning. We removed non-essential features like 'name' and post-launch variables such as 'pledged' and 'backers_count', focusing on pre-launch factors. Projects not marked as 'successful' or 'failed' were excluded to ensure clarity in our target variable. Special attention was given to filling missing values in the 'category' column. Outliers were efficiently removed using an Isolation Forest method with a contamination factor of 0.1, enhancing data precision.

In feature engineering, we introduced 'goal_usd', a standardized funding goal in USD, and addressed multicollinearity by removing highly correlated variables like 'name_len' and 'blurb_len'. We also reorganized the representation of countries within the dataset. Projects from predominant regions like the US, GB, CA, and AU were categorized into their respective groups, while the rest were aggregated into an 'Others' category. This categorization significantly streamlined our analytical process. Furthermore, we transformed categorical variables into binary format using one-hot encoding.

The final step in our preparation involved the standardization of numeric variables using a StandardScaler. This normalization ensured that our dataset was well-suited for the subsequent deployment of machine learning models, laying a solid foundation for robust and insightful analysis.

III- Section 1: Classification Model - Predicting Project Success

Three distinct classification models were evaluated: Random Forest, KNN, and Gradient Boosting. Random Forest, with its ensemble of decision trees, yielded an accuracy

of 74.57%, demonstrating its robustness in handling diverse data. The KNN model, relying on the proximity of data points, achieved an accuracy of 67.88%, suggesting potential sensitivity to the dataset's scale and variance. Gradient Boosting outperformed the others with an accuracy of 75.31%, indicating its efficacy in sequentially correcting errors from previous trees and handling the dataset's nuances.

Hyperparameter Tuning

The hyperparameters for each model were fine-tuned using GridSearchCV. For Random Forest, the optimal setup included 200 trees (n_estimators) and a tree depth of 30, balancing model complexity and learning capability and achieving a performance of 75.09%. Gradient Boosting achieved optimal performance of 75.67% with 300 stages (n_estimators), a learning rate of 0.1, and a max depth of 3, striking a balance between learning speed and overfitting. The selected model for the prediction on the grading dataset is Random Forest as it yielded a more consistent result with various random states and numerous tries.

IV- Section 2: Clustering Model - Grouping Similar Projects

In our clustering analysis, we sought to discover inherent groupings within the Kickstarter dataset, aiming to identify patterns and similarities among projects. This exploration was conducted using various clustering techniques, each offering insights into the dataset's structure. The final selected technique is K-Means Clustering with 6 clusters. Other tested methods include Hierarchical Clustering and DBSCAN but K-Means had the highest silhouette score.

To analyze the cluster results, we used feature deviation from the mean for each cluster. The results are as follows:

Cluster 1: This cluster is characterized by projects that very closely match the average project across all observed features. The projects in this cluster may represent the most typical or standard types of projects seen on Kickstarter.

Cluster 2: The slight negative deviation in both launch_at_day and deadline_day suggests that projects in this cluster tend to launch earlier and conclude earlier in the month compared to the average project. Some potential implications include early month launch and conclusion implying that creators in this cluster prefer to start and end their campaigns away from the end-of-month financial commitments that potential backers may face.

Cluster 3: Projects in this cluster are characterized by extended preparation times before going live with their campaigns. This longer interval could indicate several strategic

approaches or considerations by project creators such as careful planning, pre-launch marketing, or strategic timing to launch with a certain event or season.

Cluster 4: This cluster is markedly different from the rest in terms of goal_usd, with projects having significantly higher funding goals. This can indicate ambitious projects, possibly with a broader reach, larger-scale impact, or higher production costs.

Cluster 5: There is a small trend toward earlier hours for both launch and deadline could have several implications. Launching and ending earlier in the day could be a strategic decision to ensure that the project is live during hours of high online traffic, possibly before many people start their workday. Project creators might also be trying to capitalize on productivity and decision-making peaks, which for many people occur during the first half of the day.

Cluster 6: The noticeable deviation in launch_to_deadline_days for this cluster indicates that projects here tend to have longer campaign durations. This could suggest that creators in this cluster are giving themselves more time to reach their funding goals, which might be because they are targeting a more extensive fundraising campaign, or they anticipate needing more time to attract potential backers.

Even with slight deviations, these nuances can help in understanding the small but potentially significant differences between the clusters. For example, if projects in Cluster 5 tend to launch their campaigns during certain hours, it might suggest they are optimized for when potential backers are most active. Similarly, the small deviations seen in Cluster 2 around certain days could indicate seasonal projects or those that are timed to capitalize on certain events or holidays. These subtle trends can be valuable for targeted marketing strategies or for creators considering the timing of their campaigns. These findings also demonstrate the multifaceted nature of crowdfunding projects and the intricacies involved in grouping them into distinct categories.

V- Conclusion:

The analysis highlights the complexities and challenges in predicting Kickstarter project success. Gradient Boosting emerged as the most promising classification model, adept at handling the dataset's features and nuances. The clustering results, while not definitive, shed light on the diverse and multifaceted nature of crowdfunding projects. These insights underscore the importance of tailored feature engineering and model selection in predictive analytics, providing valuable guidance for Kickstarter project creators and platform managers.