Individual Project Summary Report

Classification Model

Choice of the Model. The model chosen is a Gradient Boosting Algorithm (GBA) which stands out as an optimal model with the highest accuracy score for predicting the success or failure of Kickstarter projects compared to decision trees, logistic regressions, and random forests. Decision trees, while intuitively interpretable, often suffer from sensitivity to changes in the training data, leading to potential overfitting. Logistic regression, on the other hand, assumes linear relationships between features and the log odds of the outcome, limiting its ability to capture complex, non-linear patterns present in the Kickstarter dataset. Lastly, GBA is preferred over Random Forest due to its sequential training, weighted observations, and efficient weight determination using gradient descent (i.e. it builds trees and corrects any 'errors' from the previous tree to improve the subsequent one). GBA uses an adaptive learning approach, focusing on misclassified instances and using weighted voting, which enables it to capture complex patterns more effectively. Therefore, it is a versatile and robust choice for business scenarios requiring improved predictive accuracy as it considers the numerous and various factors influencing them.

Feature Selection. Initially, only 'successful' or 'failure' outcomes were retained as binary variables (1 and 0). Standard preprocessing included removing irrelevant and repetitive columns (e.g., *id*, *name*, *deadline*, *created_at*, *name_len_clean*, and *blurb_len_clean*) and handling NA values. The *name_len* and *blurb_len* were instead retained for their potential impact on project success: excessively long content may deter potential backers from reading, lowering the likelihood of funding. Features post-launch were also dropped (i.e. *pledged*, *state_changed_at*, *staff_pick*, *backers_count*, *usd_pledged*, *spotlight*, *state_changed_at_* (*weekday*, *month*, *day*, *yr*, *hr*), and *launch_to_state_change_days*). Finally, *disable_communication* was removed as all

values became 'False' after eliminating other *state* values. The remaining features include *goal* (converted into USD with the *static_usd_rate*), *state*, *category* (dummified), *name_len*, *blurb_len*, *deadline_* (*month*, *day*, *yr*, *hr*), *created_at* (*month*, *day*, *yr*, *hr*), *launched_at_* (*month*, *day*, *yr*, *hr*), *create_to_launch_days*, and *launch_to_deadline_days*. After, the 'feature importance' score was explored, resulting in relatively low importance for the dummified *country*, *deadline_weekday*, *created_at_weekday*, and *launched_at_weekday* variables which were dropped, showing improvement in the accuracy score. Furthermore, based on the refined dataset, GridSearchCV, a systematic way for hyper-parameter tuning, was also applied to refine the model even more and enhance accuracy as this is one of the main objectives of the project.

Clustering Model

Choice of the Model. K-means clustering was selected for its non-hierarchical, efficient approach to partitioning data into a specified number of clusters (k). The silhouette method helped determine the optimal k as it considers both cluster cohesion and separation. The algorithm refines clusters iteratively by assigning each point to the nearest centroid, making K-means versatile and applicable to various clustering problems such as this one. On the other hand, hierarchical agglomerative clustering requires high computational power, has scalability issues, and challenges in determining optimal clusters in its dendrogram, therefore it was not chosen. Also, based on the results obtained from hierarchical agglomerative clustering with complete linkage, despite showing higher silhouette scores, the clustering with the optimal k=2 resulted in a cluster containing only one observation, making the analysis less informative and impactful. Overall, K-means is ideal in efficiency, simplicity, and versatility.

Preprocessing. A similar approach to the previous model's preprocessing was used, although all numerical variables were kept, excluding *id*. The only dropped columns in this

analysis were *id*, *name*, *pledged* (already reflected in *usd_pledged*), *disable_communication*, *currency*, *deadline*, *state_changed_at*, *created_at*, *launched_at*, *name_len_clean*, and *blurb_len_clean*. This dataset therefore aims to incorporate information pre- and post-launch to group these observations.

Insights. The optimal number of clusters using the silhouette score is 6. From observing the summary statistics in each cluster, a few notable insights regarding the average amount goals and the amount pledged were identified. The latter features were chosen as they had the most noticeable variation across the clusters. It is also worth noting that each cluster contains a relatively similar number of observations, making them comparable. So, a key insight from the clustering algorithm reveals the following: Cluster 1 represents projects with a low goal, but the highest amount pledged, indicating significant overachievement. Cluster 2 shows projects with an average goal and amount pledged. Cluster 3 involves projects with an exceptionally high goal, yet the amount pledged remains moderate. Cluster 4 comprises projects with an average or moderate goal and amount pledged. Cluster 5 features projects with the lowest goal but the second-highest amount pledged, reflecting relative success with minimal initial funding. Lastly, Cluster 6 represents projects with a moderate goal, but the lowest amount pledged. Thus, on average, projects with lower goals significantly exceed the amount they end up getting pledged and funded, which could be a strategy for project creators. In conclusion, these insights provide a nuanced understanding of Kickstarter projects, helping both project creators and backers in navigating funding outcomes and strategies. Despite similar success ratios across clusters (28% to 30%), the identified funding patterns offer valuable guidance for creators to refine strategies and backers to make more informed decisions, fostering a dynamic crowdfunding landscape.