FINAL REPORT

WHAT CONSTITUTES A SUCCESSFUL PROJECT ON KICKSTARTER?

I. Classification model

1. Data pre-processing

First, the "id" and "name" attributes are dropped as they are not relevant to the prediction model. In addition, "deadline", "state_changed_at", "created_at" and "launched_at" attributes are dropped. Additionally, attributes that can only be collected after submission are eliminated.

With the remaining attributes, "static_usd_rate" and "currency" are dropped. "disable_communication" is eliminated since they contain the same value for all the observations. "name_len" and "blurb_len" were also dropped since they correlated with the other 2 variables "name_len_clean" and "blurb_len_clean" in the dataset. In addition, all variables related to the creation of the project are also removed, as the project will only become available to the public once it is launched. Finally, "launched_at_yr" and "deadline_at_yr" are also removed as the year the past project was released would not affect the status of a new project launched in a different year.

2. Model selection

The following classification models are tested: KNN, Gradient Boosting, Artificial Neural Network, Logistics Regression and Support Vector Machine. For each model, K-fold cross-validation scores are utilized to determine the best parameter. Each model will then be run using the best parameter to determine its accuracy score. In the end, **Gradient Boosting Model with 6 samples required to split** is chosen for its highest accuracy score of 0.74 and a cross-validation score of 0.72. The model is then run with selected parameters resulting from feature selection using

Random Forest and Lasso. However, the accuracy score does not improve. Hence, the original model is retained.

3. Business implication

The classification model allows project owners to understand the positive and negative factors that impact their projects. Kickstarter could use this information to market to creators so that they could improve their chances of being successful, which in turn generates money for the site. Kickstarter could also create add-ons or applications that allow project owners to predict the state of their project and monetize the app. In addition, since the classification model contains the time, date and month that a project is launched, the project owner could choose to launch their project in a specific day in order to maximize their success rate.

II. Clustering model

1. Data pre-processing and feature selection

The features used for this model are similar to those of the classification model, with the addition of "backers_count" and "staff_pick". Since there are some outliers in "backers_count" and "goal", observations that fall within the 90% percentile of backers_count and within the 5% percentile and 95% percentile of goals would be removed.

For the clustering model, Principal Feature Analysis (PFA) is used to select relevant features. It implements a basic algorithm to select features from our dataset based on the silhouette score and *k*-means clustering. The algorithm first identifies a single feature with the best silhouette score when using *k*-means clustering. Afterward, the algorithm trains a *k*-means instance for each combination of the initially chosen feature and one of the remaining features. Next, it selects the two-feature combination with the best silhouette score. The algorithm uses this newly discovered

pair of features to find the optimal combination of these two features with one of the remaining features, and so on. The algorithm continues until it has discovered the optimal combination of n. In order to create an accurate clustering model, I choose to retain the 15 most relevant features of the dataset.

2. Model selection

For this project, I used the silhouette score to identify the number of clusters for KMeans clustering. This shows that the optimal number of clusters with the highest silhouette score is 4. The data is then clustered by using KMeans clustering with 4 clusters. Afterward, PCA is employed to reduce the dimensionality of the features to plot the clusters.

3. Business implications

Knowing the clusters that the project belongs to can help the project owner estimate their rate of success. The rate of success for each of the cluster is as follows:

Cluster	Success	Failed	Success rate
1	2288	4382	34.3%
2	31	726	4.1%
3	176	1898	8.5%
4	3	273	1.1%

As a result, Kickstarter could provide the project owner with attributes belonging to this cluster so that they could adjust the parameters of their project in order to increase their success rate. This information could evidently be monetized and thus bringing in more revenue for the site.