Amr Maraqa – 261101609 Fall 2022 Individual Project 12/12/2022

Introduction:

Kickstarter is a crowdfunding platform that offers creators a space in which they can share their vision and gain financial support for it. The incubation process of each project consists of three stages, namely creation, launch, and completion. Upon creation, creators communicate their project ideas and goals through a campaign meant to raise public awareness and interest in the project. Thereafter, a project's launch is announced, upon which the funding goal is shared and funding can be pledged to the creator until the deadline is reached. If the funding goal is attained, the project is considered a success. Otherwise, it's a failed project.

Data Preprocessing:

The Kickstarter dataset has around 15,500 projects with 45 variables each. Once the NA values were dropped, the variables known after launch were removed, anomalies detected by an isolation forest and excluded, and categorical variables dummified, the data had 11,936 observations left with 81 predictors each. The variable 'state' was separated and selected as a target variable (y), and the rest of the variables were defined as the independent variables (X).

Subsequently, X was scaled using Min-Max normalization in preparation for the feature selection process. Three feature selection techniques, LASSO, Random Forest, and PCA, were run to generate a list of predictors to be tested on the classification models developed.

Classification Model:

For comparison purposes, three types of classification models were built: K-Nearest-Neighbor (KNN), Random Forest (RF), and Gradient Boosting Trees (GBT). Each of these models underwent a cross-validation process using GridSearchCV to determine their optimal hyper-parameter combinations. The optimal combinations for each of the three models are displayed below.

KNN: n_neighbors = 75, weights = distance

	min_samples_split	min_samples_leaf	max_depth	n_estimators	max_features
RF	17	6	21	500	log2
GBT	9	3	6	250	auto

Table.1. Optimal hyper-parameter combinations for the three tested classification models

The GBT model also had its loss set to deviance and its learning rate at 0.075. Then, each model was performed on four different sets of predictors after the data into train-test sets. The accuracy scores of each of the models on the different sets of predictors are listed in the table below.

	LASSO	RF (top 10)	PCA (65 components)	All Predictors
KNN	0.6956	0.6801	0.6948	0.6903
RF	0.6852	0.7103	0.7255	0.7444
GBT	0.6847	0.6910	0.7305	0.7593

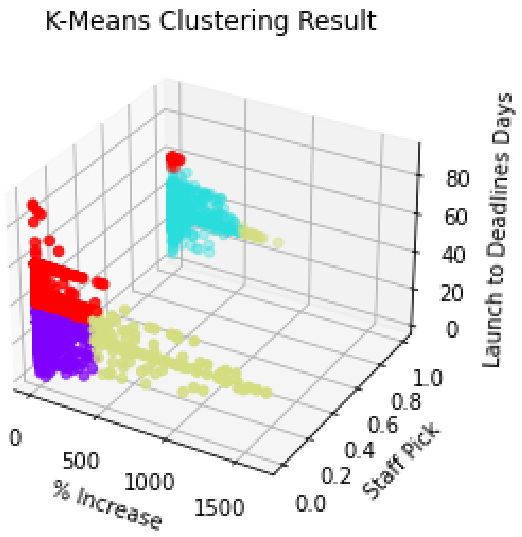
Table.2. Accuracy scores of the different classification models with four different sets of predictors

Although GBT performed better on the complete set of predictors, the RF model run on the same set of predictors was chosen as the final classification model. This decision was made due to the higher potential of interpretability the RF model provides with its feature importance ranking.

The business value this classification model holds for Kickstarter is the insights it provides into the reasons behind the success of projects. With such a model, Kickstarter can launch its in-house consulting program. From the RF feature importance list , Kickstarter can offer creators actionable advice to increase their chances of success, which consequently increases Kickstarter's revenue through elevated transaction fees.

Clustering Model:

For Kickstarter's clustering model, a similar pre-processing approach was used with respect to NA values and anomalies. This time, however, the relationship between the percent increase of the amount raised, staff-pick, and the amount of days between the launch day and the deadline day was examined. The selected method of clustering was the K-means clustering, and it was run on only the successful projects. A cross-validation process based on the silhouette score was used to determine the optimal number of clusters in the data. After running the cross-validation test over a range of clusters from 2 to 20, the highest silhouette score was obtained for 10 clusters. However, a lower number of clusters with a similar silhouette score, 4, was chosen instead for its higher cluster intrepretability. The model performance was measured using the silhouette score (0.584), Calinski-Harabasz score (3192.59) and p-value (1.11e-16), indicating that the data was well clustered. A 3D plot of the 4 clusters is shown above.



As can be seen in the result of the clustering model, the amount of money raised on top of the requested project goal is independent of whether the project was labeled as staff-pick or not. Moreover, the highest concentration of successful projects lies in the 20 to 40-day range, with percentage increase in funding increasing as the period between launch and the deadline approaches 20 days. The insights derived from this graph align well with the logic of developing a sense of urgency during a marketing campaign, and they can be utilized by Kickstarter in the consulting program suggested earlier to advise creators to prepare well before launch and target a deadline of 20 days post-launch in order to increase the probability of their project's success.