Mohamad Dalati
261113383

Task 1 - CLASSIFICATION

1 Data Preprocessing

1.1 Dropping Variables

Missing categorical values were dropped considering the large dataset; also, when replacing them with the most occurring class using impute function, the model resulted in a lower score. Irrelevant variables like *id* and *name* were dropped. In addition, predictors with words like *pledged*, *state_changed*, and *spotlight* were dropped since they were unknown at the start of the project. Moreover, variables with Dates were dropped because other variables capturing the month, day and hr exist. Lastly, *the Country* variable is related to *currency*, and after running both models separately, *currency* resulted in a lower score and was dropped.

1.2 Correlation, Outlier & Reclassification Testing

name_len and *blurb_len* were found to have high positive correlation (> 0.70) with *name_len_clean* and *blurb_len_clean* so they were dropped. Also, an Outlier test was performed, and the variable *goal* was detected. However, model performance was higher with them, so they were not dropped as they seem to represent natural variations. Lastly, since none of the categorical variables had a very high number of classes(>50), reclassification was not required.

1.3 Dummifying Variables

Categorical variables like *disable_communication, country, category, deadline_weekday, created_at_weekday,* and *launched_at_weekday* were transformed into numerical indicators.

2 Model Selection

For the best accuracy score with a reasonable interpretation level and minimal overfitting, GBT was selected. Upon splitting the model, random_state was set to 600 for reproducibility purposes. For optimal parameter selection, hyper-parameter tuning was performed using GridSearchCV.

3 Model Results

Dataset was split into 77% training set and 33% testing set for serving predictions. The model scored 75.70% for accuracy, 65.83% for precision, 52.03% for recall, and 58.12% for F1 score. The most important predictors are *goal, category_Web, create_to_launch_days, launch_to_deadline_days, name_len_clean,* and *category_software.*

4 Business Insights

The model proposes the following recommendations for Kickstarter project creators, use the project budgeting tool available on Kickstarter to calculate the project funding goal, as it is essential. Focus on projects related to *Web* and *software* categories because they are more trending than others. Moreover, set the number of days between the creation and launching of the project to its deadline after extensive research since it is critical. Lastly, keep the project name without words like "for" or "and" and make it as concise as possible!

Task 2 - CLUSTERING

1 Data Preprocessing

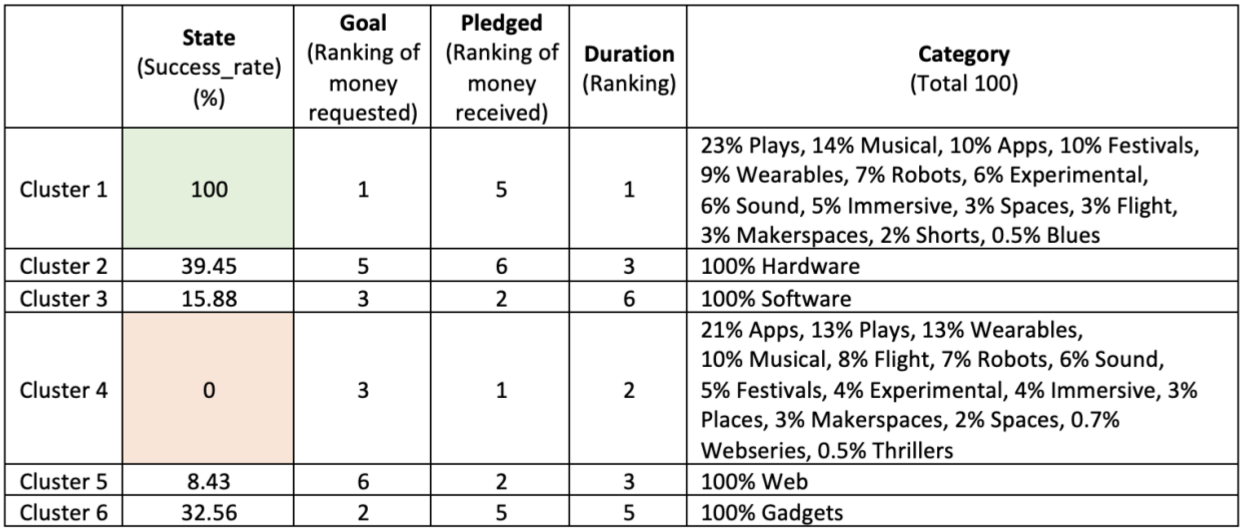
Similar processes and checks were performed. However, only five variables that could bring value were selected. They include *state, goal, pledged, launch_to_deadline_days,* and *category.* Also, unlike GBT, clustering variables need to be standardized to reduce the variance, and due to the high variance between variables, MinMax Normalization was performed.

2 Model Selection

After comparing the Elbow test and Silhouette score to find the optimal number of K, Non-Hierarchical K-Means clustering with k = 6 was selected for this model. Since the Elbow test only minimizes the distance within clusters, the Silhouette score was considered.

3 Model Results

Ranking Score: 1 Lowest - 6 Highest (Compared to other clusters)



	State (Success_rate) (%)	**Goal** (Ranking of money requested)	**Pledged** (Ranking of money received)	**Duration** (Ranking)	**Category** (Total 100)
Cluster 1	100	1	5	1	23% Plays, 14% Musical, 10% Apps, 10% Festivals, 9% Wearables, 7% Robots, 6% Experimental, 6% Sound, 5% Immersive, 3% Spaces, 3% Flight, 3% Makerspaces, 2% Shorts, 0.5% Blues
Cluster 2	39.45	5	6	3	100% Hardware
Cluster 3	15.88	3	2	6	100% Software
Cluster 4	0	3	1	2	21% Apps, 13% Plays, 13% Wearables, 10% Musical, 8% Flight, 7% Robots, 6% Sound, 5% Festivals, 4% Experimental, 4% Immersive, 3% Places, 3% Makerspaces, 2% Spaces, 0.7% Webseries, 0.5% Thrillers
Cluster 5	8.43	6	2	3	100% Web
Cluster 6	32.56	2	5	5	100% Gadgets

Based on the "Results" df at the end of the code attached, cluster center interpretations are provided in the table above. For instance, Cluster 2 contains projects from the hardware category with a success rate of 39%. Compared to other clusters, they have the second-highest money request and receive the highest amount of money, with an average duration rate.

4 Business Insights

The model recommends Kickstarter project owners/creators the following: for a project to be successful, consider asking for the lowest amount of money, setting the duration of the project from launch to deadline to be short, and consider creating projects in categories like Plays, Musical, and Apps and avoiding categories like Web, Webseries, and Thrillers. On the other side, for Kickstarter project creators interested in Hardware category projects, ask for higher goal amount, since there is a high success chance of receiving more money than requested upon setting their project duration at an average compared to other projects. For project owners in the Software domain, avoid placing the project duration too long! this will inevitably lead to the project's failure. Unlike project creators in the Gadgets domain, if you want your project to be successful, set it at long durations. Lastly, avoid launching projects in categories like Blues, Shorts, and Spaces, as they seem to have less interest or attention than others.