**Individual Project**

Elaine Yu - 261030551

McGill University, INSY 446: Data Mining for Business Analytics

Professor Warut Khern-am-nuai

December 4, 2022

## Introduction

Kickstarter is a crowdfunding platform allowing creators to bring innovative projects to life. In this report, we are interested in building a classification model using the Kickstarter dataset to predict whether the state of a project is prosperous or failed and a clustering model to group similar project characteristics.

## Classification Model

### Data Preprocessing

We execute the classification model when the project owner submits the project. Hence, before building it, we remove invalid features, specifically variables that cannot be observed at the time of prediction. After importing the file into Spyder, we first dropped any observations containing “live”, “cancelled”, and “suspended” from the variable “state” so that our target variable only has “successful” and “failed” as values. After then, we removed invalid features due to the reasoning above, including any variables about deadline, launch, state change, pledged, spotlight, and backers. We also discarded the project ID and name since they were not valuable for our prediction. Finally, null values are dropped.

Next, categorical variables including “state”, “disable_communication”, “country”, “currency”, “staff_pick”, “category”, “created_at_weekday”, “created_at_month”, “created_at_day”, “created_at_yr”, “created_at_hr” are dummified.

### *Feature Selection*

To remove even more irrelevant data and reduce the complexity of the model, we used Random Forest for feature selection. LASSO and PCA are not used in this project as our dataset contains many categorical variables. Since these two modelling methods require all predictors to be on the same scale, standardizing these features could lead to difficulties, and the model would perform poorly.

From the feature selection process, we chose the following variables to be our predictors since they had the highest feature importance: “goal”, “staff_pick”, “name_len”, “name_len_clean”, “blurb_len”, “blurb_len_clean”, “create_to_launch_days”, and “category”.

## Develop Classification Model

The model we have chosen for this project is Random Forest due to its low risk of overfitting, unlike the decision tree, and good efficiency and accuracy. In addition, we are dealing with many categorical predictors, and logistic regression tends to perform better with numeric data. If we look at the codes in the Python file, the random forest has better scores than logistic regression. With random forest, the accuracy score is 76.44%, which is ideal and realistic. The precision score of the model is 67.64%, with a recall score of 55.65% and an f1 score of 60.87%. This means that if the algorithm identifies a project as successful, the probability of the algorithm being right is 67.64%. Among successful projects, the likelyhood that the algorithm would identify them as successful is around 55.65%.

## Clustering Model

In general, clustering algorithms do not perform well with categorical data since it is discrete and does not have a natural origin. Some people use K-Modes for mixed data, but we have not covered this concept in this class. Hence, we will focus on K-Means in this project.

## Chosen Variables

The chosen variables are the following: “state_successful”, “goal”, “name_len_clean”, “name_len”, “create_to_launch_days”, “blurb_len”, and “blurb_len_clean”. These variables are numerical data, except “state”, which is dummified.

## Develop Clustering Model

First, using the Elbow Method, we found that the optimal number of clusters for our K-Means algorithm was 2. With this new knowledge, we used the silhouette method to evaluate cluster cohesion and separation. The average silhouette score for Cluster 0 is 0.6383 and 0.6143 for Cluster 1. Hence, the average silhouette score is 0.6222, higher than 0.5. This number shows that the model provides good evidence of the reality of the clusters in the data. By running K-Means, projects in the first cluster tend to have a state marked as successful, a higher goal amount, a longer length in the number of days between project creation and the public launch date, and a longer length in the project name and project blurb. Conversely, projects in the second cluster are similar in terms of having the status “failed” and lower amount and length in the selected variables, as mentioned in the previous sentence.

## Conclusion

In conclusion, we built a classification model using Random Forest to predict if a project is deemed as successful or failed. We found that the accuracy score of the model is around 76.44%, which is ideal and realistic. In addition, we built a clustering model using K-means that has a silhouette score of 0.6222, which illustrates good evidence of the reality of the clusters in the data.