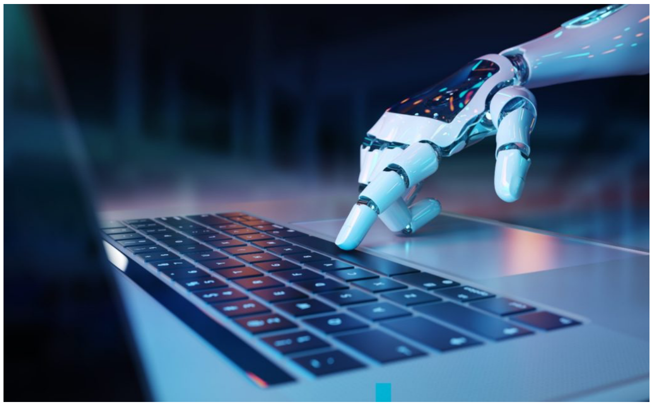
DESAUTELS | McGill



INDIVIDUAL PROJECT

Kickstarter Classification and Clustering Model Analysis

Mesaye Bahiru 260634934

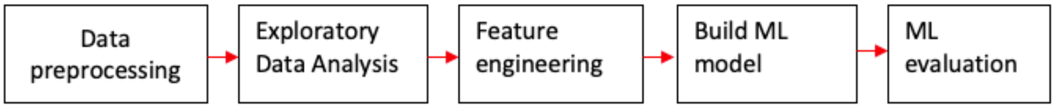
Introduction

Kickstarter is crowdfunding platform; its mission is to help people bring their projects to life. It is place for creative, innovative, and game changing ideas across different categories to flourish

Main objective of the project: the goal of this project is using classification model to predict if project will be successful or failure at the launch and perform unsupervised learning via clustering model to study pattern or partition data into non-overlapping clusters.

Kickstarter Classification model

Classification model is suitable when the target variable is categorical data (in this case *state* is success or failure). The flowchart to go through for a successful classification model.



Data preprocessing

Data preprocessing needs to be performed on the Kickstarter dataset to reduce poor machine learning model. The *state* column was filtered into successful and failed, because the analysis was interested to predict whether the project will success or fail. The data type for column *deadline*, *state_changed_at*, *created_at, launched_at* changed into datetime object. Subsequently, the duration it took from the *creation to launch* was calculated. The isnull () sum was used to check null values and it was observed *category* and *launch_to_state_change_days* contain null values. Then, using dropna function category was dropped. Lastly, check if there were any duplicates in the dataset.

Exploratory Data Analysis (EDA)

EDA provides an opportunity to visualize and understand the dataset more effectively. The column *goal* contains amount in local currency, and this was converted to USD currency for consistency.

the time takes from creation to launch was visually examined, and most projects get created and launched right way. Number of campaigns launched per year was checked and the data shows consistent increase in projects launched and month of November and July experience higher project launched. USA has highest projects contribution by country.

Feature Engineering

The target variable is column *state*, categorical variable using NumPy *where* function it was converted into numerical value. Then, success rate was checked by *country* to see where most successful and less successful projects are located. In both cases, USA comes first, likewise, it was better to modify the country data into USA and others. The *category* column was checked to see which category has the top success rate; hardware scored the highest. Using, *get_dummies* function country and category data converted into numerical values. The final engineered dataset was ready to be placed into machine learning model.

Machine Learning Model

Logistic regression was used to build the model. The process was presented in the code, the goal is to have deep understanding on how well the model reacts with different predictors and examine the accuracy. The procedure follows define x and y variables, split train/test set, model build and predict. Then, roc curve was produced to visualize the accuracy of the different models and the highest accuracy score achieved 72 % but the F1-score was 37%. This was not satisfactory. Hyperparameter tuning of logistic regression using *gridsearchcv* was used to improve the logistic regression model and with this analysis 48% F1 score was achieved. Additional analysis was done using random forest. define x and y variables, split train/test set, model build and predict. Random forest produced 70 % accuracy and 52% F1 score.

Classification model for Business Context

Three groups (project creator, Kickstarter platform, backers) will benefit from this analysis. The data shows close to 67% of the projects fail, likewise, this model provides insight into which category to choose (hardware, gadgets) shorter duration for prep/ campaign, and these insights help newcomers on the platform organize their project plan and backers also have better understanding of which project to back based on this analysis.

Clustering Model

Principal analysis component (PCA) was performed to determine number of components needed to explain variance with cut-off threshold of 95%, and 23 number of components satisfies the variance. For clustering model, K-means algorithm was used. Elbow method was utilized to determine the optimal k value (the optimal cluster was 4). Then define x values in this case *pledged*, *backers_count,* followed by standardize method. Build and fit the kmeans model with k=4 and visualize the clusters. There were four clusters, in cluster 1 projects with few pledged amounts see limited backers. Cluster 2 as the number of pledged amounts increase the backers count stays constant. Cluster 3 showed significant increase in pledge amount attracts higher backer counts; it takes lot more pledged amount to get more buzz that induces more backers. Interestingly, there was also few high pledged amounts, while the number of backers was very low. There were selective investors who wants specific projects. Cluster 5 showed the opposite of cluster 1, projects with highly pledged amount related to higher backer count.

Clustering model for Business Context

Clustering pattern presented the relationship between backers and pledged amount, with this model Kickstarter platform has better understanding of the investors behaviour and how money is allocated into projects. Using this information, the platform can attract more investors and maintains existing investors by promoting high quality projects that meets the demand.