Data Mining for Business Analytics

Individual Project

April 15, 2023

Elina Gu

McGill | DESAUTELS

Overview

The goal of this individual project is to choose techniques/algorithms to build two models on the Kickstarter dataset. A 3 double-spaced pages report along with the Python codes, Classificaiton.py and Cluersting.py are submitted to explain details and project insights.

1. Classification Model

Develop a classification model (i.e., a supervised-learning model where the target variable is a categorical variable) to predict whether the variable state will take the value successful or failed. The classification task is assumed to be executed at the time each project is submitted to Kickstarter. In other words, at the moment the project owner submits the project, the classification model will predict whether the project is going to be successful or not.

2. Clustering Model

Develop a clustering model (i.e., an unsupervised-learning model which can group observa- tions together) to group projects together and explain the characteristics that you observe in each cluster.

Classification Model – Gradient Boosting Tree

The classification model aims to predict the success or failure of Kickstarter projects based on project properties at the time of creation. The data pre-processing involved dropping rows with values other than "successful" or "failed" in the "state" column and empty rows in predictor variables (X). The following variables were selected for the classification model:

1. Goal: The requested amount plays an essential role in determining the project's success. A goal set too high could make it difficult to achieve success.

2. Category: The category in which a project falls influences its success. Some categories are more popular than others;
3. Project_len: This calculated column, determined by subtracting the created_at year from the deadline year, determines the project's total length. Longer project lengths are more likely to succeed.
4. Name_len_clean and blurb_len_clean: Cleaned lengths of the key words in the project name and blurb are critical factors influencing the project's success. The project name can attract potential backers and influence the project's success.
5. Static_usd_rate: This is an indicator of the origin country. The Static_usd_rate is fixed throughout the project's duration, and it can potentially influence/change the amount requested.
6. Created_at_yr, created_at_hr, created_at_weekday: The year, hour, and weekday of a project's creation influence its success. The year could affect economic climate or trends, and the hour and weekday could affect the project's visibility to potential backers.
7. Deadline_yr, deadline_hr, deadline_weekday: The year, hour, and weekday of a project deadline determine a project's success. The deadline creates a sense of urgency for potential backers, influencing the project's success.

Predictors were standardized using the standard scaler and split with a test size of 0.33. The training dataset was used to create a Gradient Boosting Tree model without tuning any hyperparameters, resulting in an accuracy score of 0.7455. Next, cross-validation was used to determine optimal values for hyperparameters including subsample, n_estimators, min_samples_split, min_samples_leaf, and max_depth for the reasons of avoid over-fitting and simplicity. The final accuracy for the model is 0.7467.

Clustering Model - DBSCAN

The DBSCAN clustering model was utilized to group similar Kickstarter projects based on their characteristics. To accomplish this, two variables were created: "region" and "category," to group projects by their geographic location and category. The "region" variable differentiated between projects from North America and the rest of the world, while the "category" variable segmented projects into five areas: Entertainment, Academic and Learning, Technology, Music and Sound, and Physical Space. The clustering model was built using these variables, along with "goal" and "created_at_year." Prior to the analysis, the location and category groups were dummified and X is standardized. After various trials, the DBSCAN clustering model was finalized with an eps value of 3.2 and a minimum sample size of 20, producing 10 clusters (with 24 noise points) and a silhouette score of 0.6978.

The largest cluster, Cluster 0, consisted of 5,794 projects with a goal of $51,868, created in 2015 and originating from North America. Projects in this cluster were mostly in the Technology category, demonstrating that technology-related projects are favoured.
Cluster 1 included 1,930 projects with a goal of $50,167, all created in 2015 and originating from North America, mostly in the Entertainment category, showing that Kickstarter has a strong presence in North America.

Cluster 2 contained 744 projects with the third-highest goal of $58,314, created in 2014 and originating from the rest of the world, mostly in the Technology category.

Cluster 3 had 541 projects with a goal of around $44,484, created in 2015 and originating from North America, primarily in the Music and Sound category.

Cluster 4 consisted of 273 projects with goals of $37,919, created in 2015 from North America, primarily in the Physical Space categories. Cluster 5 had 245 projects with goals of around $40,548, created in 2015 from the rest of the world, mostly in the Entertainment category.

Cluster 6 included 244 projects with goals of around $39,930, created in 2014 from North America, primarily in the Academic and Learning categories. Cluster 7 had 203 projects with the highest goals of around $64,919, created in 2014 and originating from the rest of the world in the Physical Space categories. Cluster 8 consisted of 78 projects with the lowest goals of around $34,187, created in 2015 and originating from North America, primarily in the Academic and Learning categories.Cluster 9 had 77 projects with the second-highest goals of around $58,947, created in 2015 and originating from the rest of the world, mostly in the Music and Sound category.

Insights: North America is the primary origin of the projects in most of the clusters. This suggests that Kickstarter has a strong presence in North America. The technology category groups are present in two of the top three clusters, followed by Entertainment, indicating their popularity in Kickstarter projects. Projects in the Academic and Learning and Music and Sound category groups are present in several clusters but are less common overall.