Jonathan Gaudaire (260966733)

INSY466 - Data mining for business analytics

Individual project

I. Classification task

The objective of the classification is to predict whether a Kickstarter project will be successful at the time the project is submitted. To do this, I will use the Kickstarter dataset provided, which is composed of 18568 observations and 45 variables.

1. Pre-processing

The first step is to drop observations with the state variable not equal to either “successful” or “failed”. We will then set 1 for successful observations and 0 for failed ones. As the prediction is done at the time the project is submitted, I also dropped variables that would be known at that time. These variables are ‘pledged’, ‘staff_pick’, ‘backers_count’, ‘usd_pledged’, and ‘spotlight’, as well as all variables related to state change or launch. I also dropped the variable ‘disable_communication’ as all values were equal to 0.

I also created dummy variables for the categorical predictors, such as ‘category’ and ‘country’.

2. Feature engineering

Using variables available in the dataset, it is possible to create new predictors. Thus, using the static_usd_rate variable, I added the variable “goal_usd” to convert all amounts into USD. Moreover, I defined ‘diff_deadline_created’ as the number of days between the deadline and the created_at date.

3. Independent variables

The independent variables used are goal, name_len, blurb_len, blurb_len_clean, deandline_month, deadline_yr, deadline_hr, created_at_month, created_at_yr, country, category, goal_usd, and diff_created_deadline. The use of these variables slightly increases the accuracy of the models run, so I included all of them. Other variables decreased the accuracy of the models, therefore I excluded them from the independent variables used.

4. Model selection

To select which model to use for the classification, I decided to compare the accuracy scores of Logistic regression, Decision tree, Random forest, Artificial neural networks, and KNN for multiple hyperparameters and multiple feature selection combinations. The highest accuracy score was found using Random Forest with max_features = 6. (Accuracy score = 0.7582 tested on the Kickstarter y_test dataset).

5. Business use

This model can help Kickstarter to determine which projects are more likely to be successful as soon as they appear on the website. Using this, they could improve their “project recommendation” page, as well as the “staff pick” and “spotlight” flags. It can also help users to better understand the chances of success for projects they want to back.

II. Clustering algorithm

1. Model selection

To perform clustering on the Kickstarter dataset, I decided to use Kmeans. Indeed, when dealing with many observations, it is better to avoid hierarchical clustering.

2. Feature selection

When building a clustering model, feature selection is crucial. Here, my objective is to cluster Kickstarter projects based on the variables present in the Kickstarter dataset provided. To obtain useful insights into those projects, I will cluster them based on a small number of features from which I can gain some insight. The features selected are goal_usd, backers_count, pledged_usd, staff_pick, and state. This way, I can segment the projects based on these variables.

3. Number of clusters

To choose the appropriate number of clusters, I computed the silhouette scores for models with n_clusters between 2 and 10 using the features mentioned above. As the clusters were all well apart and distinguished from each other for n_clusters = [2,3,4,5,6,7,8,9,10] (silhouette score > 0.7), I decided to look at the output for each kmeans model. The description of the clusters is provided below.

4. Cluster description

- Cluster 1 (Non-ambitious successful staff-picked projects): This cluster is composed of projects that were highlighted by the website's staff, as staff_pick = 1 for all observations. The mean goal was $40913.9, which is way lower than the average goal for all other clusters except cluster 3. 80% of the time, these projects were successful.
- Cluster 2 (Non-ambitious unsuccessful non-staffed-picked projects): None of the projects belonging to this cluster were successful. In the same way, none of them were "staff_picked". The average usd_goal was slightly higher than for cluster 1.
- Cluster 3 (Non-ambitious successful non-staff-picked projects): All these projects were successful even if they were not staff-picked. The average goal was only $13196.2, which is the lowest average goal.
- Cluster 4 (Ambitious unsuccessful non-staff-picked projects): All these projects were unsuccessful, but their usd_goal was way higher. However, the average amount pledged is also almost 0. These projects almost all raised $0.
- Cluster 5 (Successful ambitious staff-picked projects): These projects are all successful, and they were staff-picked.
- Cluster 6 (Unsuccessful ambitious non-staff-picked projects): these projects are similar to projects belonging to cluster 4, but they still raised $62246.3 on average.

This segmentation can be used by the users to better understand whether their project is going to be successful. It can also be useful to choose the projects that should be staff-picked.