**Predicting the success of a crowdfunding campaign**

In this project, we delve into the realm of crowdfunding and focus specifically on Kickstarter, a popular platform for creators to bring their ideas to life. Our objective is to develop a classification system that can accurately predict the success or failure of projects on Kickstarter. By analyzing the crawled database, we aim to provide valuable insights to guide creators in setting up their project campaigns effectively and help them make informed decisions on launching Kickstarter campaigns based on their project topics.

The dataset we have gathered from Kaggle.com and Webrobots.io serves as the foundation of our investigation. Through this report, we will try to introduce, describe, visualize, and manipulate the dataset to extract meaningful information.

Procedure and first steps.

Initially, we acquired the data from Kaggle and then proceeded to present it in the form of a DataFrame.

The process for obtaining the data from Webrobots.io was more complex. We encountered two types of files, namely CSV and JSON. After careful consideration and multiple attempts, we made the decision to only extract the data from Webrobots.io for the month of December 2020. For data prior to December 2020, we relied on the previously mentioned data from Kaggle.

Our objective was to merge the two DataFrames from Kaggle and Webrobots. However, we discovered that the "Category" variable differed between the two datasets. In the Kaggle dataset, "Category" was split into two variables, namely "Main\_category" and "Sub\_category". On the other hand, the Webrobots dataset contained the "Category" value concealed within a dictionary structure. The dictionary included information similar to the Kaggle dataset, encompassing Main and Sub\_categories. As an initial approach, we opted to utilize the main category to facilitate the merging of the two DataFrames.

Afterwards, we conducted an extensive examination of the variables within each DataFrame and finalized our selection of the following shared variables:

Data Classification:

* [Data Description](https://docs.google.com/spreadsheets/u/0/d/1BWL2sR6jcQO4UXzMpfv_rfsTejqkao7wapFjhAXRYgM/edit)
* [Interpretation of the Data features](https://docs.google.com/document/u/0/d/1ZyP122bMtJ4t8UK_lpgFv6ddLcrXnxDTzMKTcKSbZ1s/edit)
* Launched\_at: Represents the date on which each project was launched.
* State: Signifies the current status of each campaign, providing valuable insights into its outcome and success.
* Country: Indicates the geographic location or country associated with each project.
* Goal: Reflects the funding target in USD for each campaign, enabling businesses to assess funding strategies and evaluate project feasibility.
* Pledged: Represents the actual amount of money pledged for each campaign in USD, serving as an important indicator of campaign support.
* Category: Identifies the main category of each project, offering an overview of the project landscape and facilitating market analysis.
* Backer\_count: Denotes the number of individuals who have pledged their support to a specific project, serving as an indication of the project's popularity and success.

Upon finalizing the variable selection, we proceeded to merge the two DataFrames, incorporating the chosen variables into a new consolidated DataFrame.

In order to determine the relative success of different projects, we began by exploring the following questions:

* Which category exhibits the highest number of projects?
* Which category garners the greatest number of backers?

Next we are moving on and testing chart distributions in order to find interesting insights from our data.

These steps provide an overview of the initial procedures undertaken in the project.

—------------------------------------------------------------------------------17/05/23