Identification of sustainability-focused campaigns on the kickstarter crowdfunding platform using NLP and ML boosted with swarm intelligence

Data Analysis: part 1

Submitted by: Jossin Antony

Affiliation: THU Ulm

Date: 11.06.2024

Overview

Introduction

The aim of the project is to study how crowdfunding campaigns support sustainable inititatives. This project, in particular, focuses on crowdfunded campaigns in the kickstarter platform and explores a dataset of c.a 184,186 initiatives from different domains (e.g, Technology, Music, Publishing etc.). The goal of the analyses here is to find the most important features that are relevant to initiatives that are both sustainable as well as profitable. The analyses will also explore the possible relationship of the features with each other, and elucidate insights that might contribute to better understanding of the success/failure propsects of current and future environment focused crowdfunded initiatives.

Details of dataset:

1. Source: Kickstarter_File.xlsx
2. Generation mode: provided by researcher
3. Time period considered: 04-2009 to 05-2021 (c.a 146 months).
4. Total entries: 184,185

The initial data preparation consists of examining the various features and eliminating redundant features & renaming and re-ordering of features and saving the dataframe.

Preparation of Dataset

First we make sure the dataset is 'reasonable', i.e, it has good structure, columns have data of expected types, devoid of null values etc.

The basic information of the data is as following:

```
The dataframe has 184187 rows and 24 columns.

The overall dataframe information is given below:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 184187 entries, 0 to 184186
Data columns (total 24 columns):
 #   Column                    Non-Null Count   Dtype  
---  ------                    --------------   -----  
 0   blurb                     184184 non-null  object 
 1   Environmental             2053 non-null    object 
 2   Social                    2053 non-null    object 
 3   state                     184186 non-null  object 
 4   Subcategory               184186 non-null  object 
 5   Unnamed: 5                176465 non-null  object 
 6   converted_pledged_amount  184186 non-null  float64
 7   country                   184186 non-null  object 
 8   country_displayable_name  184186 non-null  object 
 9   created_at                184186 non-null  object 
 10  currency                  184186 non-null  object 
 11  deadline                  184186 non-null  object 
 12  fx_rate                   184186 non-null  float64
 13  goal                      184186 non-null  float64
 14  launched_at               184186 non-null  object 
 15  duration                  184186 non-null  float64
 16  name                      184186 non-null  object 
 17  pledged                   184186 non-null  float64
 18  slug                      184186 non-null  object 
 19  staff_pick                184186 non-null  float64
 20  state.1                   184186 non-null  object 
 21  static_usd_rate           184186 non-null  float64
 22  usd_exchange_rate         184186 non-null  float64
 23  usd_pledged               184186 non-null  float64
dtypes: float64(9), object(15)
memory usage: 33.7+ MB
None

We also make the preliminary observation that the columns named 'environmental', 'social' and
'unnamed: 5' have lots of 'NaN' values. We will deal with them later.
```

Next we provide meaningful names to the columns to reflect the nature of the data they contain as well as re-order them.

```
'The new column_names are:'
['campaign_name',
 'blurb',
 'slug',
 'main_category',
 'sub_category',
 'is_environmental',
 'is_social',
 'country',
 'country_displayable_name',
 'created_at',
 'launched_at',
 'deadline',
 'duration_in_days',
 'currency',
 'goal_in_local_currency',
 'pledged_in_local_currency',
 'usd_pledged',
 'pledged_amount_usd',
 'staff_pick',
 'state.1',
 'fx_rate',
 'static_usd_rate',
 'usd_exchange_rate',
 'is_success']
```

Next we drop the columns which are redundant or which do not add any value to the analysis. The dropped columns are as following:

1. **'country' and 'country_displayable_name':**

   We need only one of these; but we save the country codes for later reference.

2. **'created_at', 'launched_at', 'deadline', 'duration':**

   There is no discernible difference between 'created_at' and 'launched_at' since they are, at maximum, only few days apart in order to have an effect on the results we look for. 'duration' provides the difference in days between launched_at and deadline and we keep this parameter (for now).

3. **'currency', 'goal_in_local_currency', 'pledged_in_local_currency', 'usd_pledged','converted_pledged_amount_usd', 'fx_rate', 'static_usd_rate', 'usd_exchange_rate':**

   There is the goal- but only in local currency- and the pledged amount- in both local currency and usd. We add a new column, 'goal_in_usd', which gives the goal in usd as well. It is obtained by multiplying the 'goal_in_local_currency' with the provided 'usd_exchange_rate' (Logic: The converted_pledged_amount_usd is provided by the author as a product of 'usd_exchange_rate' and 'pledged_in_local_currency').

4. **'staff_pick' and 'state.1':** These columns are dropped, since state.1 is a reptition of the column 'is_success' and 'staff_pick' do not seem to add value to the analysis at hand.

5. **'slug' and 'campaign_name':** 'slug'is a repetition of 'campaign_name', it is dropped.

```
Overview of the selected features:
 campaign_name
blurb
main_category
sub_category
is_environmental
is_social
country
duration_in_days
goal_usd
pledged_amount_usd
is_success
```

We also drop the rows which have 'NaN' values in more than 3 columns.

Note: We expect 'NaN' values in atleast 2 columns, is_envt and is_social.

Next we remove those rows which do not have the 'campaign_name','blurb', 'main_category', 'sub_category', in the expected string format.

We also remove the rows where the columns 'duration_in_days','goal_usd', 'pledged_amount_usd' also do not have data in the expected number format.

We also strip spaces from 'main_category', 'sub_category' and 'country' columns.

We replace the duration in days with duration in months, rounded to the nearest month.

```
duration_in_months
1    136022
2     29238
0     10861
3       341
4         1
Name: count, dtype: int64
```

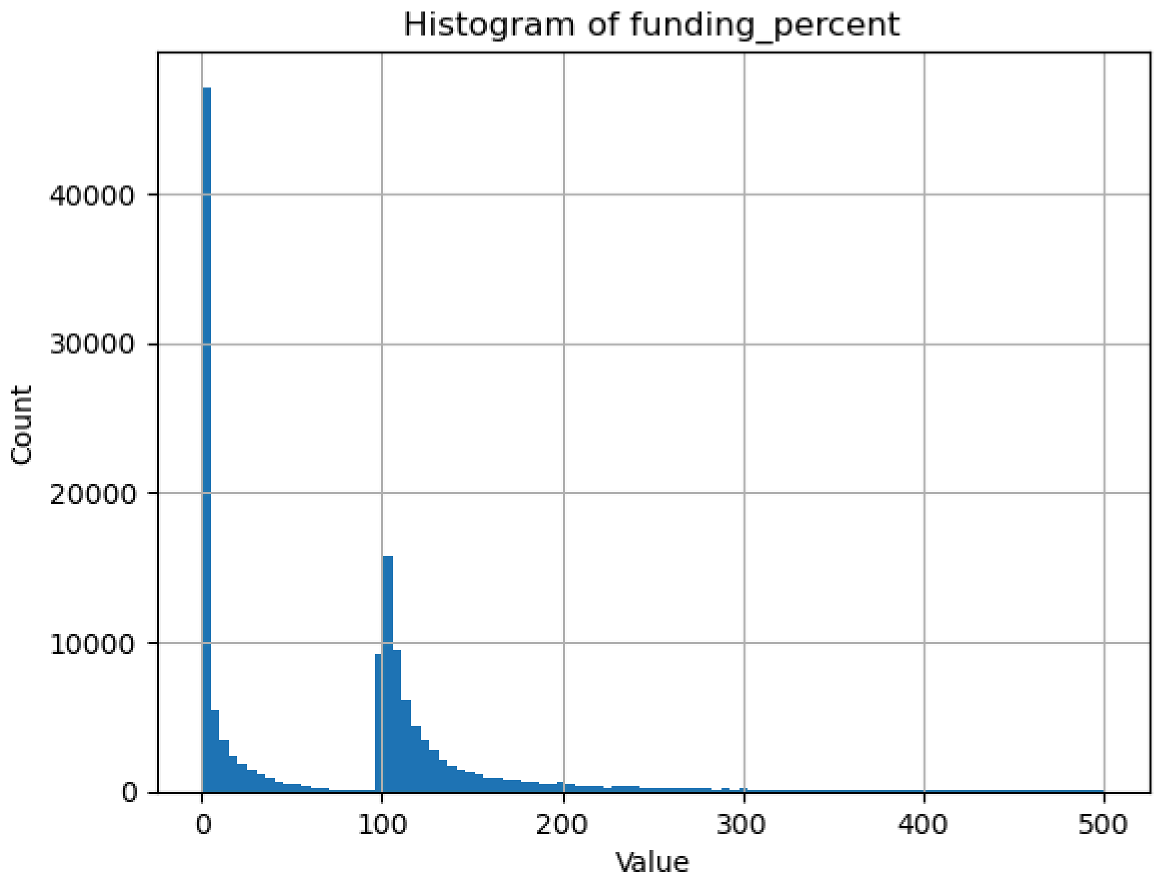
We observe that the values of the 'duration' are reasonable within the scope of the project.

Important:

We consider campaigns where the goal is atleast USD 1000.

We now observe how much percentage of the funding were acquired by the campaigns.

We now observe the Histogram.

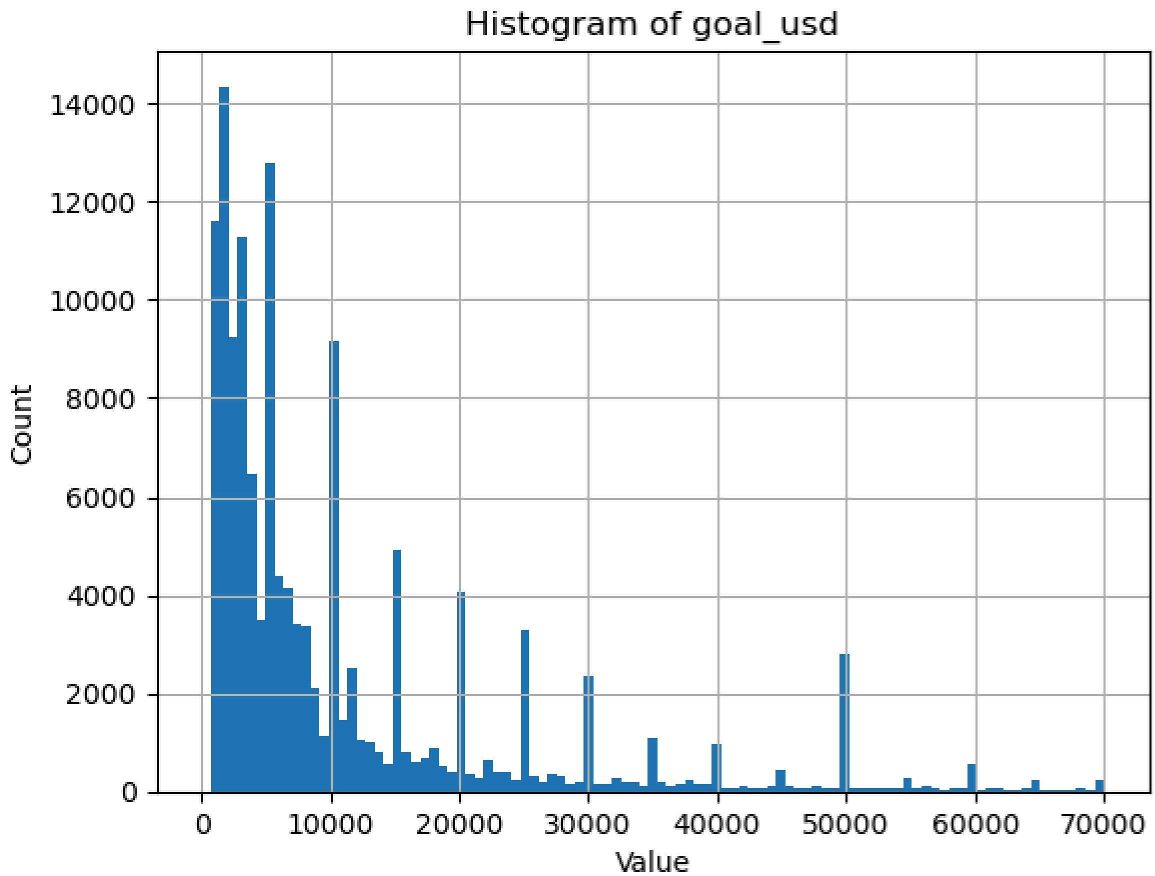


From the histogram, we discern that there can be 3 different categories for the is_success status. These are:

- funding_percent < 100%: failure
- 100% <= funding_percent < 300%: success
- funding_percent >= 300%: blockbuster

```
is_success
goal_achieved    70357
fail             66720
blockbuster       8779
Name: count, dtype: int64
```

We also categorize the campaign goal based on the amount. Perhaps, we observe later trends with respect to goal amounts and the chances of campaigns being successful.



The histogram shows that the goal amounts are centered on multiples of 5000s. The majority of the campaigns aim to raise an amount less than USD 50,000. The follwing categories are defined:

- 1000 <= goal_usd < 10,000: 1k-10k
- 10,000 <= goal_usd < 50,000: 10k-50k
- goal_usd > 50,000: 50k_plus

We now drop the 'goal_usd', 'pledged_amount_usd' and 'funding_percent' columns since the information in these columns are captureed in the 'goal_usd_category' and 'is_success' columns respectively.

We also consider only those rows which are in english language. We first remove rows whose descriptions do not appear in latin script. Further more, we delete rows which contain only links. We also try to deduce the language of the description and retain only those rows whose description is provided in english. Note:

- The rows containing only links for description are found using a url_regex. It is not perfect and cannot detect all rows with urls only. We resort to manualk deletion of such rows (In our case: 1 row only).
- The detection of languages is implemented by the langdetect package. It also provide false negatives. But since these are negligeble compared to the total data corpus, we disregard the false negatives.

Now we check if there are any unprocessed rows. If yes, detect the language in these rows.

```
Processed samples: 145791
Empty/Unprocessed values: 0
Samples with non-english descriptions: 6204
```

We remove the rows whose descriptions are not in english.

```
The final dataset has 139587 samples.
```

```
The columns in the dataset are:
 campaign_name
blurb
main_category
sub_category
is_environmental
is_social
country
duration_in_months
goal_usd_category
is_success
```

There are 'NaN' values for the 'is_envt' and 'is_social' categories. We will populate them in the next set of analyses.

Save dataset

After this we save the data to a local file for the next set of analyses.