UNIVERSITÀ DI PISA

Master Degree in Data Science and Business Informatics

Distributed Data Analysis and Mining project

Kickstarter Projects Analysis

Andrea Napolitano 667031 (a.napolitano12@studenti.unipi.it)

Eva Manai 257280 (25728026@studenti.unipi.it)

Michele Velardita 578770 (m.velardita@studenti.unipi.it)

Paolo Andriani 579604 (p.andriani@studenti.unipi.it)

Steffania Sierra Galvis 663718 (s.sierragalvis@studenti.unipi.it)

Academic year 2023-2024

Table of contents

1 Introduction

For our study we choose a part of Kickstarter Projects dataset, retrieved on Kaggle platform at Kaggle - Kickstarter Project. The dataset contains 375000 records about crowdfunding worldwide projects data from 2010 to 2018; it has 15 features and an imbalance rate of 35% on positive class, namely the success of a project. Our goal is to discover those key characteristics that increase the probability of success of a campaign and those factors that most influence investors. To do this, efforts were concentrated on a multivariate analysis of the data in order to discover possible correlations and trends, and a predictive analysis, training classification and regression models and extrapolating their feature importance. Given the large size of the dataset, the Apache Spark framework was used to conduct the analyses.
In the following, we will detail the steps we followed in the main phases of our investigation, starting with data exploration and extraction to suit our purposes. Subsequently, we will address various analyzes on the data, including multivariate, textual analyses, and observations on the feature importance of classification and regression models, and then conclude with an exposition of the results obtained and the conclusions we have drawn.

2 Data exploration and preparation

From the original set of features (names in parenthesis) we see that data is organized into 15 main areas of interest *(main_category)* as movies, food, book, games, etc..., and further divided into more specific 159 subcategories *(categories)*. Records are collected across 9 years *(year)* from 23 different countries *(country)* using 14 different currencies *(currency)* to measure collected funds *(pledged)* and the money amount requested to complete the project *(goal)*, within the starting *(launch)* and ending *(deadline)* dates. Also, the number of donors *(backers)* is stored, and a conversion to US Dollars of both pledged and goal values *(usd pledged, USD_pledged_real, USD_goal_real)*. The first one converted by Kickstarter platform, and the other two by using fixer.io API to convert currencies. Each project has an identification number *(ID)* and a status value (*state*) that can be success, failed, canceled, live, suspended or undefined.
From this original feature set we notice the need to improve the informative potential of data described by exploiting inner information in a more efficient way, also suitable for models we'd like to run later. As first step of data preparation we first select only records whose state is successful or failed, dropping all rows of other kind of status because we want to study a binary classification task and it is not possible to translate meaningfully these other status into the selected two. Then, we delete *id, goal, pledged, usd pledged* columns and rename *usd_pledged_real* into *pledged* and *usd_goal_real* into *goal*, to have two unique monetary measures with which compare records. Depending on which model we use later on, we take other preparation steps, described, if present, in dedicated paragraphs.

2.1 Feature extraction

Given the limited presence of descriptive variables, new ones were extracted from the already existing features. In particular, the year, month, day of launch and the time interval between the start and end of a project (expressed in days) were extracted from the deadline and launch date. From the title, the length and a binary categorical variable were extracted; this last variable (*use_of_"?!"*) tracks the use of symbols within the title. The *country* column has several values with a very low frequency. Therefore it was decided to group the countries by continent, in order to have a more balanced distribution and more usable information. Table 1 summarizes the features extracted from the dataset.

Original Feature	Extracted Features
Launched	year month day_of_the_week
Launched + Deadline	time_interval
Title	length use_of_"?!"
Country	continent

Table 1: Features extracted from existing ones.

2.2 Multivariate Analysis

Considering exclusively the successful and failed projects, the different distributions and correlations of the variables were analysed.

In our initial analysis, we delved into a detailed examination of the continent where projects originated, as illustrated in Figure 1. America prominently emerges as the frontrunner, closely followed by Europe. However, it's crucial to acknowledge that these statistics may not offer a comprehensive reflection of the global crowdfunding landscape. The data we've utilized is specific to the Kickstarter platform, providing valuable insights into trends within the Americas and Europe. Nevertheless, it overlooks crowdfunding activities on platforms in Asia and other regions. Given that different continents host their own popular crowdfunding platforms, it's important to interpret these findings with the awareness that they might not capture the entirety of the global crowdfunding scenario.

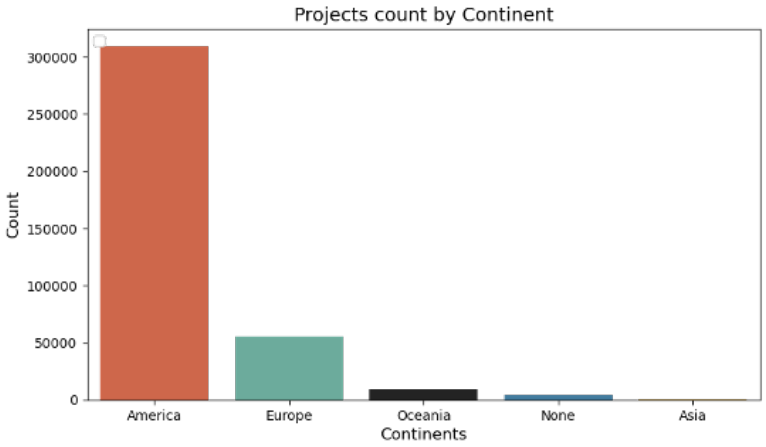


Figure 1: Distribution of started projects over continents

After this simple analysis, how the relationship between success and failure varies over time was explored. Figure 2 shows how most of the projects were launched after 2014, due to the fame the platform achieved in those years.

It is possible to notice a trend in the relationship between failed and successful projects, as projects with a positive outcome tend to decrease compared to others in the years in which the platform had the greatest following. Instead, analyzing the distribution taking into account the months, a slight imbalance is highlighted towards the failed projects in the summer months of July and August. The length of time between the start and finish of a project, taken alone, does not appear to be directly related to the outcome; in fact, the average of the time intervals of the campaigns with positive and negative results is very similar (32 and 35 days respectively). However, analyzing more in details the relationship between the time interval of a campaign and the success rate, it

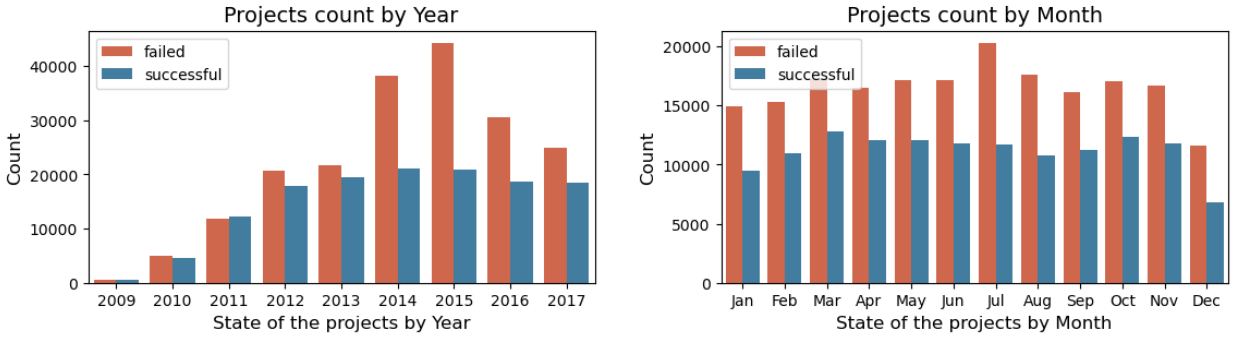


Figure 2: Distribution of successful and failed projects over years and months

can be noticed that a shorter project duration can slightly increase the probability of success of a campaign.

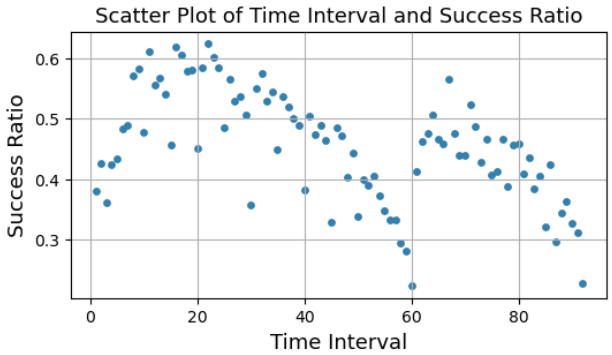


Figure 3: Relationship between the success rate and duration of a campaing

Furthermore, it was also analyzed how the day of the week on which a campaign is started has a causal relationship with the outcome; the day a project launches should obviously not affect the status of a project.

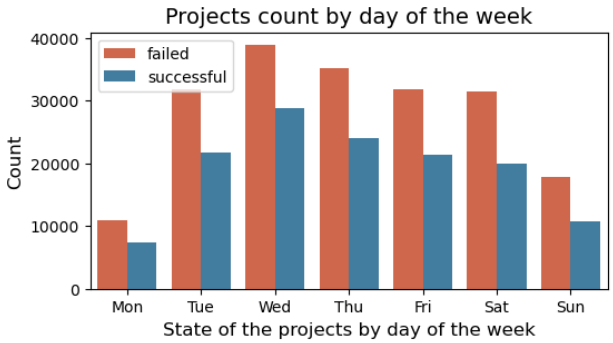


Figure 4: Distribution of successful and failed projects over days of the week

Although few projects are launched between Sunday and Monday (Figure 4), the ratio remains more or less constant. The variation in title length between successful and unsuccessful campaigns shows a slight discrepancy, as the latter seem to have a slightly shorter length, as shown in Figure 5.

The use of exclamation and question marks is rare in titles. Despite this, it seems to slightly

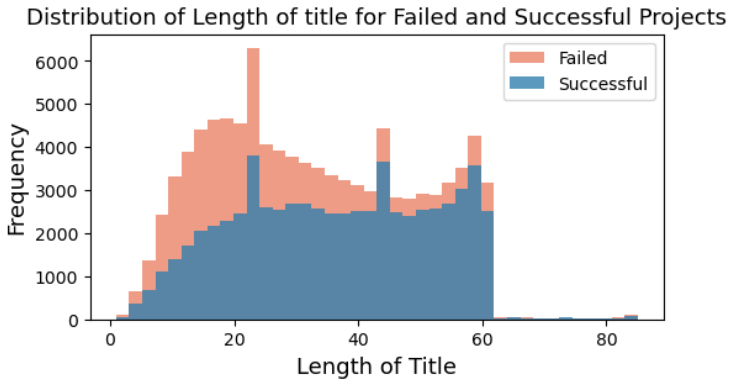


Figure 5: Analysis of length of title

influence the outcome of a campaign. Figure 6 shows how the relationship between success and failure is much more balanced among those projects where symbols are used.

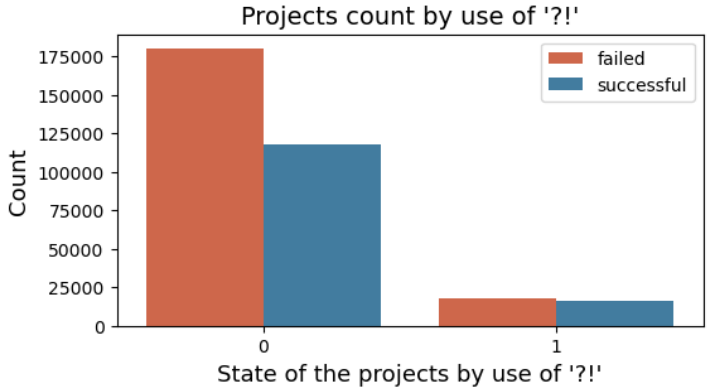


Figure 6: Analysis of usage of symbols

Afterwards, the number of supporters and the goal set for each campaign were analyzed (expressed in dollars). To be able to visually examine them, both columns were transformed using the logarithm; in fact, in the dataset there are several projects with a number of backers equal to 0 and several projects with a goal close to 0.

Overall, the goal of a campaign does not allow us to separate the two classes well, and both follow the same distribution (Figure 7). However it is possible to notice how projects with really big funding goals often fail, probably because they're very risky, and it's hard to get enough people to support such huge campaigns.

It was attempted to answer the question whether a higher number of backers contribute to the success of a project or not. By plotting the distribution of the number of backers separately for campaigns with positive and negative results, we notice a clear separation between the two; the Figure 8 illustrates a clear pattern: successful projects often attract a large number of backers. This occurs because substantial backer support enhances a project's likelihood of success.

In the dataset there are two features that group the projects: *category* and *main_category*. The former can be considered sub-categories with a higher granularity, while the latter are less detailed macro categories, with 159 unique values for *category* and 15 unique values for *main_category*. For the data exploration phase, macro categories were mainly taken into consideration; furthermore,

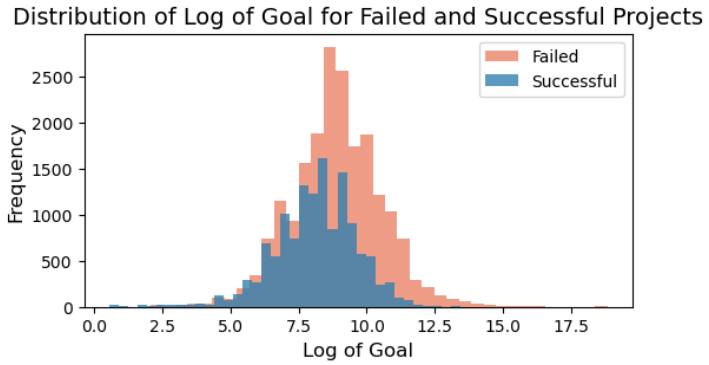


Figure 7: Funding goal of failed and successful projects

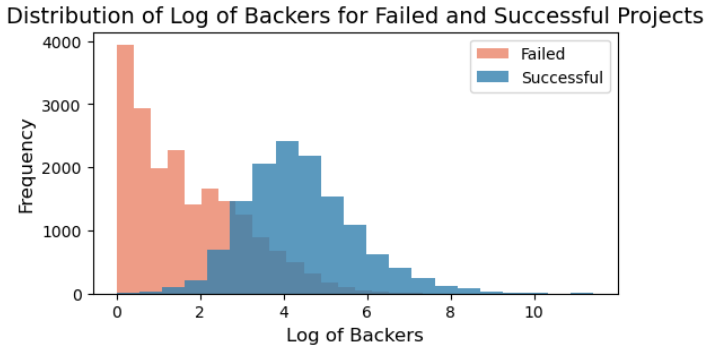


Figure 8: Number of supporters of failed and successful projects

as an evaluation metric, the ratio between the number of successful campaigns and the number of failed campaigns was considered. The most frequently published types of projects are *film and video, music*, and *publishing*. Analyzing the most frequent ones it was discovered that musical projects are very successful. Film and Video projects, however, don't always make it; they have a high probability of both failing and succeeding. This shows us that different types of projects within the same category can have very different results. Tables 2 and 3 summarize the characteristics of the best 3 categories and the worst 3 categories according to the metric mentioned above.

Rank	**Category**	**Success/Fail Ratio**	**Frequency**
1	Dance	1.9	3573
2	Theater	1.8	10242
3	Comics	1.4	9878

Table 2: Top 3 best categories for success/fail ratio

Although the Dance category is among the most niche ones, it is, at the same time, the one that attracts the most supporters in order to reach the set goal and therefore to be categorized as successful. The technology category, one of the most popular among those who start campaigns, is also the one that has the worst chance of success.

To explore these phenomena in more detail we decided to also take into consideration the time factor, and observe how certain trends can change over the years.

Categories with a higher success rate seem to more or less follow the same trend. Figure 9 shows the best 2 categories, divided by success and failure, as the years pass.

Rank	Category	Success/Fail Ratio	Frequency
1	Technology	0.31	27050
2	Journalism	0.32	4149
3	Crafts	0.37	7818

Table 3: Top 3 worst categories for success/fail ratio

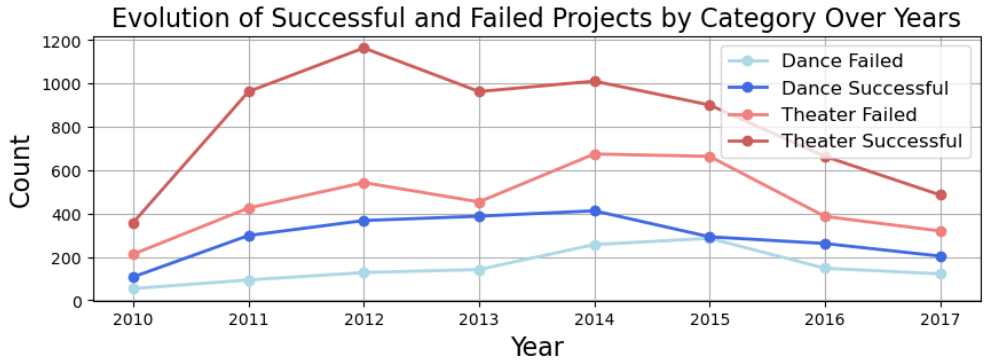


Figure 9: Trend of the top 2 best categories over years (the darker lines represent successful projects, while the lighter lines represent failures).

Among the categories explored, the technological one seems to follow a more singular path: there is a strong increase, with a peak around 2015, in the failure of this type of campaign (Figure 10). The trend related to technological projects can be influenced by several factors: After 2014, Kickstarter and other crowdfunding platforms have become increasingly popular, attracting a wide range of projects and, at the same time, the increase in the number of projects could lead to greater competition and market saturation; in a context where there are many similar proposals, it may be more difficult for some projects to stand out and attract the attention needed to achieve the desired funding.

2.3 Project title analysis

An in-depth analysis was conducted on the project titles to answer the question whether some words (or sequences of words) are more present in successful projects and whether they can influence the success of a campaign. All stopwords have been removed from the titles, and everything has been

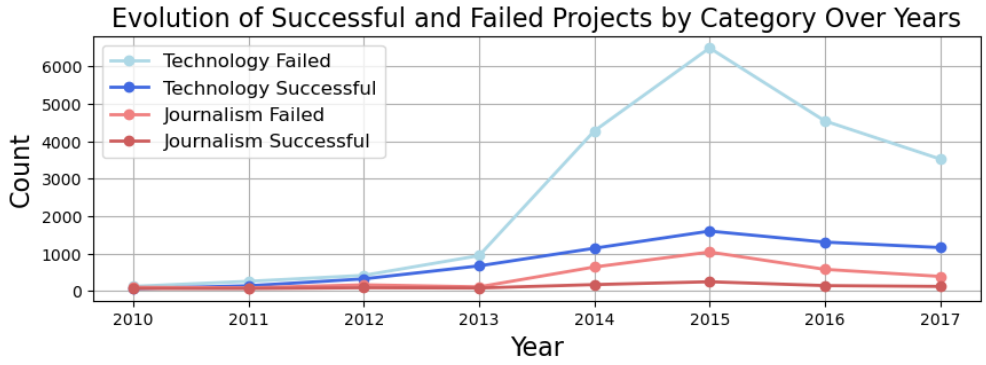


Figure 10: Trend of the top 2 worst categories over years (the darker lines represent successful projects, while the lighter lines represent failures).

changed to lower case. Through a word cloud, a visual overview of the most used words was extracted. Only the most frequent words in the titles of successful projects and failed projects are represented, where the size of the words is proportional to their frequency. This allowed us to quickly identify the most relevant terms divided by campaign outcome. From Figure 11 it is possible to notice several similarities between the two sets, such as the strong presence of terms such as "project", "game" and "book".

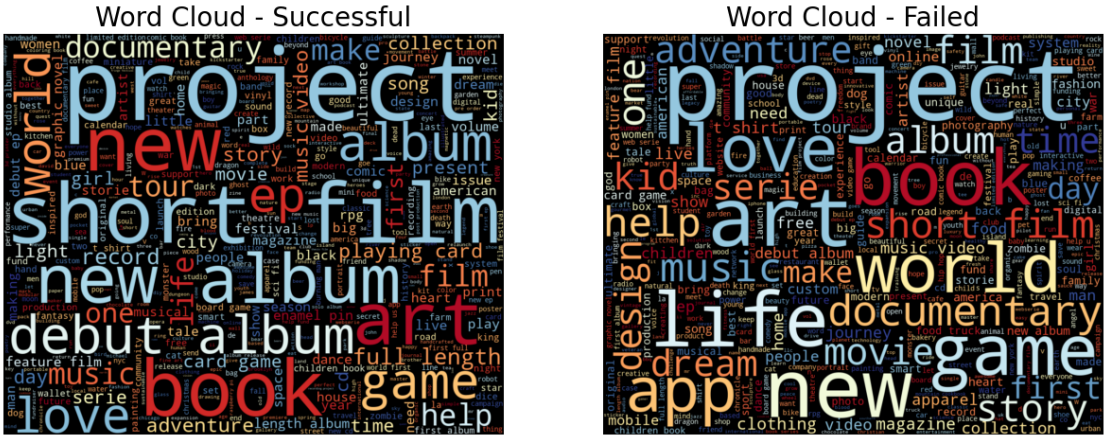


Figure 11: Word Clouds of more frequent words in successful and failed project's titles

Analyzing individual words there are several overlaps between the titles of successful and failed projects as shown in Figure 12.

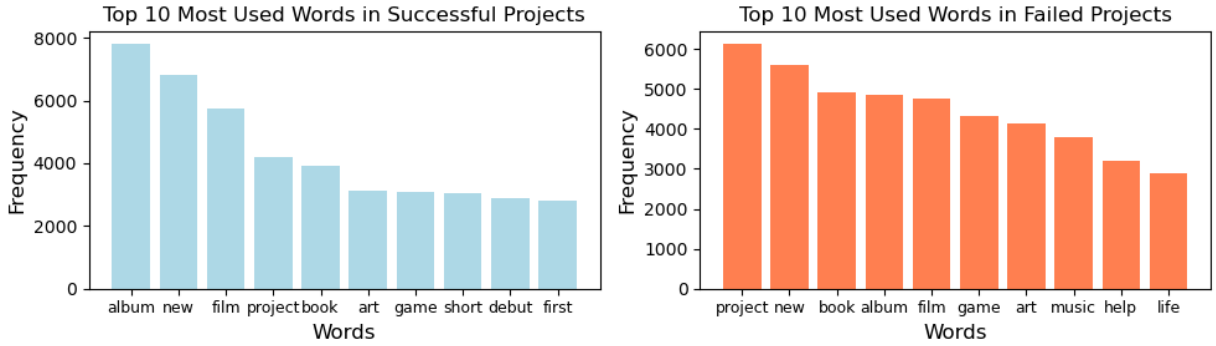


Figure 12: Top 10 most used words in successful and failed projects

2.4 Bigrams analysis

A similar analysis to the previous one was conducted by extrapolating the top bigrams of the projects titles. The bigrams are the most common pair of consecutive words of the titles. As before, all the stopwords were removed so the analysis was done only on the meaningful words of the title. In Figure 13 we can see the top 10 bigrams for successful and failed project. Notice that there are some similarity within the two groups and some similarities too with the top 10 words shown in Figure 12. It is interesting notice that even if the result are similar, the frequencies between the two group are different, successful bigrams have a greater count than the failed one, even considering that successful project are the minority ones. For highlighting this difference, the percentage of failed and success, calculated dividing the number of successful bigrams by the total count of that bigram. The result obteined are reported in table 4

Bigram	Success %	Fail %
new album	70.8	29.2
playing card	65.8	34.2
debut album	65.4	34.6
short film	65.1	34.9

Table 4: Success/failure percentage of top bigrams

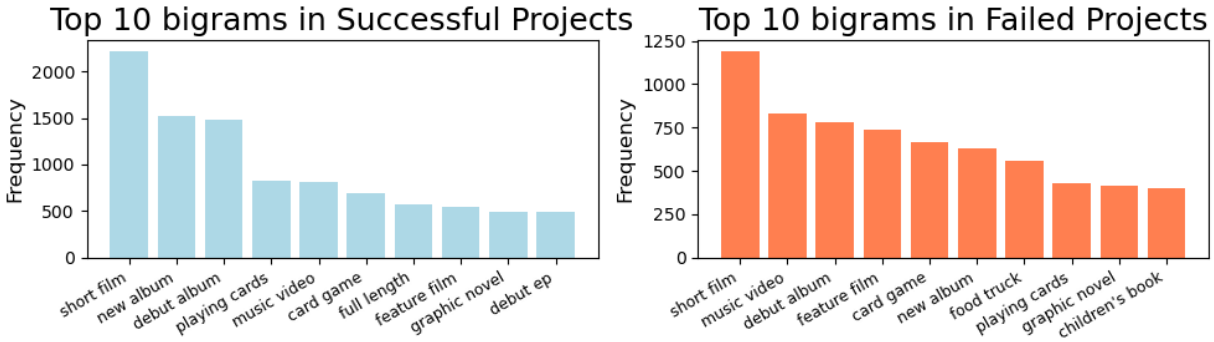


Figure 13: Top 10 bigrams in successful and failed projects

2.5 Correlation analysis

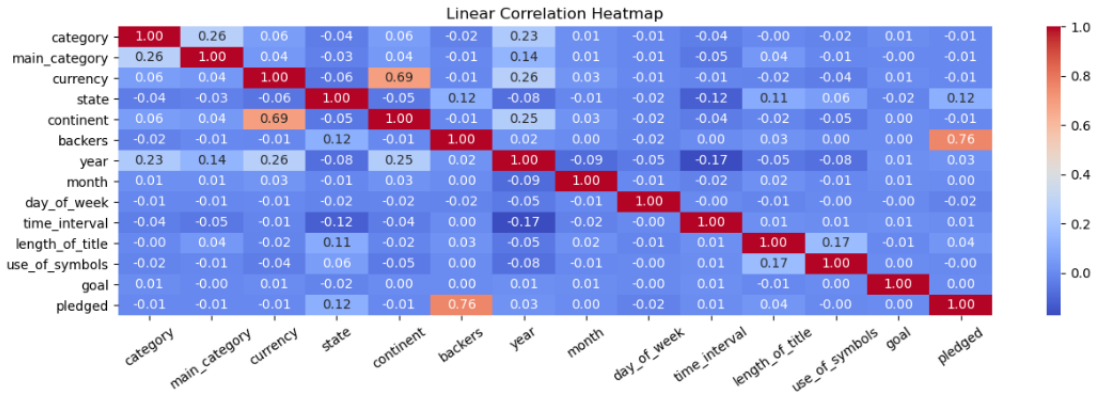


Figure 14: Pearson coefficient between all the variables

After the previous exploration of the data, a correlation analysis was performed to see if we could find some other relation between the variables. First, it was calculated the Pearson coefficient between all the variables. This calculation gave as result the heat map in Figure 14

The correlation analysis revealed a significant positive correlation between continent and currency, and between *backers* and *pledged*, indicating a moderate to strong relationship. This finding is consistent with the meaning of the variables, because, the higher the number of donors, the greater the amount collected. Less obvious relationships revealed in the analysis are between *year* and *category*, and *year* and *currency*. This could be due to social or economic events that have occurred in certain years in different countries.

As the result of the linear correlation analysis was not so satisfactory, it makes sense to do the analysis with the Spearman coefficient. In this case the results are shown in the heat map in Figure 15. The relations between *continent* and *currency*, and *backers* and *pledged* also appears

in this analysis, but with a much higher value. As well, new relations can be seen. For instance, *state* and *pledged*, and *state* and *backers* show a moderate to strong relationship. However, this result does not add something new to our analysis, because their relation is intuitive. New weaker relationships appeared, for instance in *goal* and *time interval*, and in *pledged* and *length_of_title* have a positive correlation coefficient of 0.22 and 0.17 respectively. Those relations are less intuitive than the previous ones, and could be considered in further analysis.

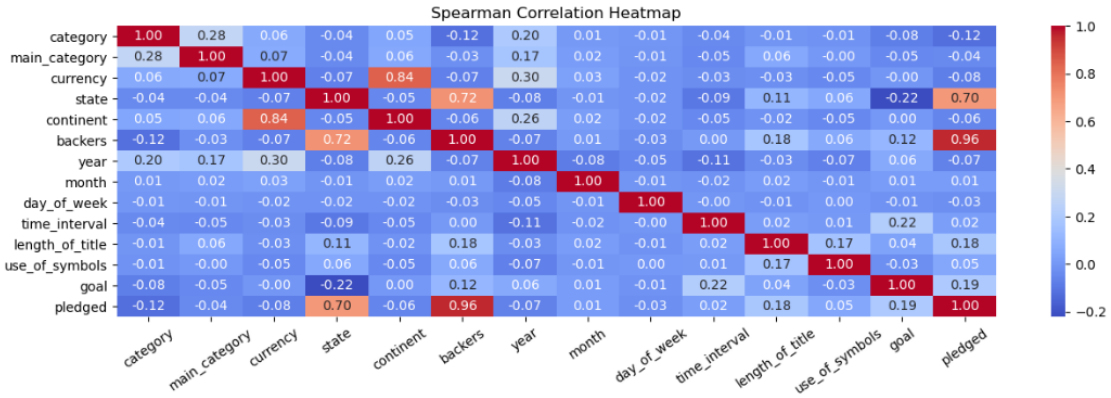
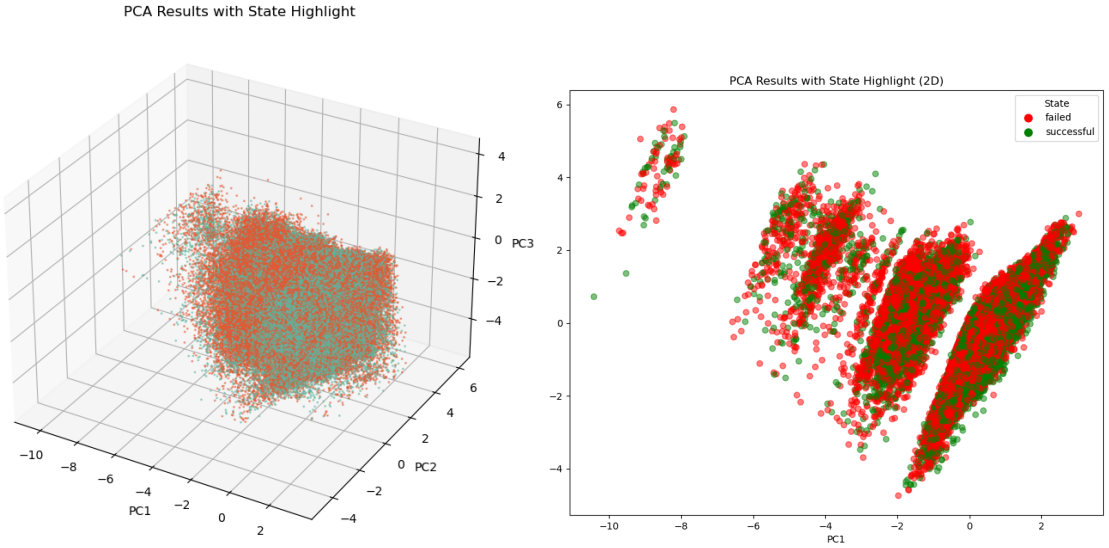


Figure 15: Spearman coefficient between all the variables

2.6 Principal Component Analysis (PCA)

We made some analysis by the usage of the Principal Component Analysis (PCA) projecting the following features: *goal, year, month, day_of_week, time_interval,length_of_title, use of "?!", Category, Main Category, Currency, Continent.* At start, three components has been used, presenting challenges in terms of interpretability, as we can see in figure 16a, due to its intricate nature. Notably, the color scheme aids in distinguishing project outcomes: red represents failed projects, while green indicates successful ones. However, due to the inherent complexity of the three-dimensional representation, it is difficult to identify specific patterns, and, importantly, success and failure points are not distinctly separated within the PCA analysis.



(a) (b)

Figure 16: Principal Component Analysis of (a) 3 Components, (b) 2 components

Running PCA with two components yields a remarkably clear depiction of distinct clusters within the data, as we can see in figure 16b. These well-separated clusters suggest the potential influence of a categorical variable, contributing to the formation of discernible patterns. Considering this observation, further analysis could involve the application of a clustering algorithm to assign labels to these clusters and gain insights into the common characteristics that unite them. By applying a clustering algorithm, it becomes possible to uncover inherent patterns and relationships within the data, shedding light on the factors that contribute to the formation of these identifiable clusters.

3 Classification

Our main objective is to predict whether a project will be successful or not. To do that, it was decided to drop some feature that were considered useless or too much informative. For example, we dropped the *pledged* feature that is an aposteriori information that can bias our classification models, certainly, if is greater than the goal then the project is successful, so using it carried out perfect classifiers that did not provide useful information of the data. We noticed that even the *backers* feature carried out too good results, it is trivial that a project with a lot of backers will be successful, so it was discarded for the classification task as well.

The dataset has been split into train and test with a ratio of $70\% - 30\%$. Due to the strong presence of outliers in goals and backers, the standard scaler produces a distorted view of the original distribution. For this reason it was decided to use a robust scaler, which is less sensitive to extreme values. In order to apply the classification models, the data was preprocessed via the following pipeline:

- String Indexer on categorical columns (including *year, month* and *day of the week*) and on target column *state*.
- Vector Assembler on numerical columns and normalize them using Robust Scaler.
- Vector Assembler to merge in a single vector categorical and scaled numerical columns.

3.1 Decision tree

A decision tree was trained to perform the classification task regarding the *state* variable. A grid search with cross validation (4 folds) was performed in order to find the best hyperparameter for the tree, the metric used for the validation was the accuracy. The parameters tested were: maxDepth: (10, 15, 20), minInstancesPerNode: (60, 80, 100),that is the minimum number of instances in a child leaf for consider the split valid, and the two impurity measure (Gini, Entropy). The best parameters found were maxDepth = 15, minInstancesPerNode = 80, and entropy as imputiry measure. Using the best tree obtained the resulting measure on the test set were the following

Accuracy	**Precision**	**Recall**	**F1**	**AUCROC**
0.674	0.613	0.518	0.562	0.646

Table 5: Measure recap of the DT

3.1.1 Random forest

Some test with a random forest classifier was done but, due to the complexity of the model and the time required for training, it was not feasible doing an exhaustive grid search to find out the best parameters; only two hyper-parameters were tested (numTrees: [30,40,50] and maxDepth:

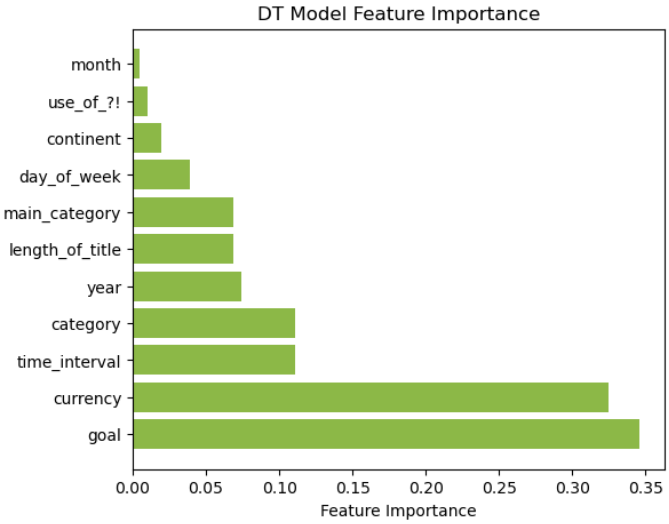


Figure 17: Feature importance of Decision Tree

[2,3,4]) resulting in poorer performance in comparison to the decision tree. Table 6 shows the performances achieved with the model.

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.639	0.658	0.234	0.346	0.575

Table 6: Measure recap of the random forest model

3.2 Neural Network

A grid search was performed, with respect to the number of hidden layers and learning rate, with cross validation (4 folds), trying to maximize the area under the ROC curve. As results was obtained a model with a hidden layer structure of 6, 4, 2 nodes, and a step size of 0.01. The performances achieved were poor, obtaining an AUROC of 0.67 and an f1-measure of 0.62. Subsequently, we tried to obtain the same performances with a simpler neural network (hidden layers = 2, 2, 2 nodes). The Table 7 summarizes the structures and performances achieved by the two neural networks.

hidden layers	**step size**	**AUROC**	**F1**	**Accuracy**
(6, 4, 2)	0.01	0.676	0.627	0.640
(2, 2, 2)	0.05	0.671	0.622	0.638

Table 7: MultiLayer perceptron performances with different hyper-parameters on test set

Due to the lack of flexibility of the multilayer perceptron and the impossibility of using other activation functions other than the Sigmoid, the neural network achieved lower performance than other classification models.

4 Regression

4.1 Logistic regression

The logistic regression model was applied to the dataset for predicting the outcome. However, the achieved performance did not reach high levels as we can see in the Table 8. Despite the term "regression," the chosen evaluation metrics resemble those commonly used for classifiers, as the model's output is a binary label.

Accuracy	**F1**	**AUROC**
0.613	0.572	0.642

Table 8: Measure recap of the Logistic Regression

4.2 Linear regression

Since the linear correlation analysis showed a high correlation between *continent* and *currency*, and between *pledged* and *backers*, we need to choose only two of them to avoid multicollinearity in our model. Because the *currency* variable has a finer granularity than *continent*, we prefer to remove *continent* from the model. Considering that *pledged* could say more about the state of the project than *backers*, this last one won't be consider for the model. Notice that in this case we are considering only the Pearson coefficients obtained, and not the Spearman ones. This is because, we will train a linear regressor. However, in the next subsection, the variables *pledged* and *backers* both won't be consider because they have a high positive Spearman coefficient.

To train a multi-linear regressor, the categorical variables, and the numerical discrete variables like *month*, *year*, *day_of_week*, and *use_of_symbols* were encoded using the one-hot encoder. Those variables needed it to be encoded because their labels acts only like labels and the order is not important. So, to avoid that the model interprets the index assigned to every category as ordinal, the one-hot encoder is need it.

The linear regression model was trained with 70 percent of the original data, and the other 30 percent was used for testing the model. As expected from the correlation analysis, the linear regressor model did not performed too well. The root mean square error was of 0.36, and the R-squared score was 0.48. The last one means that 48% of the variance of the state is explained by the other variables. If we analyse both metrics together, we notice that the high rmse value, and the small r^2 value indicate that the model is not making accurate predictions, and it is not explaining much of the variability in the dependent variable.

Despite of the performance, let's check what the model says about the feature importance. Figure 18 indicates that the category and goal variables play an important role when trying to predict the state of the project.

There is a parameter of the linear regressor that has not been used yet. This is the regularization technique. To see if the performance of the model improves when using one of these techniques, we train again the linear regressor using the elastic net regularization, that combines both Lasso and Ridge regularizations. Additionally, for this model the pledged variable will be eliminated.

The new model gave the following rmse and r^2 values on the test data: 0.49, and 0.004 respectively. As can be seen, these new considerations did not improve the result of our model; on the contrary, it seems to worsen it.

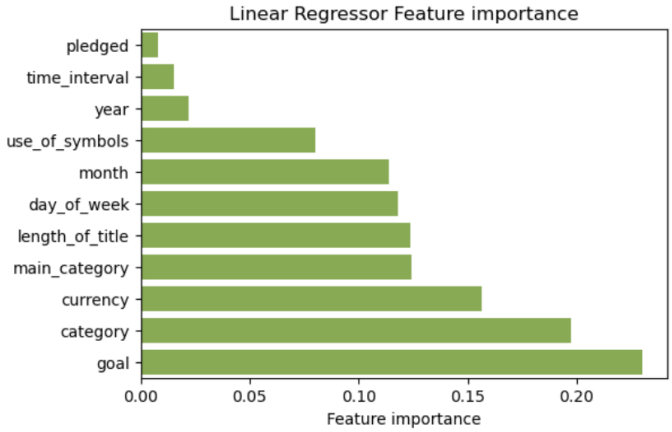


Figure 18: Feature importance as revealed by the linear regressor.

4.3 Gradient Boosting

A grid search was performed to tune some GBT classifier hyperparameters using cross validation with 4 folds, and trying to maximize model's AUROC. As result (with values maxDepth = 5, maxIter = 30, and stepSize = 0.2), it was achieved slightly better results than the ones reached with multilayer perceptron model, but again not optimal ones: two models were trained, changing some categorical features passed as input to the classifier. The first attempt (that involves the categorical features *main_category, currency, year, month, day_of_week, continent, use_of_?!*), achieved an AUROC value of 0.732, and an F1-score of 0.673. As second attempt, it was decided to remove *currency* feature and insert *category*, the theme of the projects with an higher granularity. Very similar performances were achieved (with a slighlty improvement), summarized in the Table 9.

Attempt	**AUROC**	**F1**	**Accuracy**
first	0.732	0.673	0.680
second	0.753	0.688	0.693

Table 9: Gradient Boosting Machine performances on test set with different features considered

Feature importance was also explored, to understand which values the model is mostly based on and which attributes can influence the outcome of a campaign more than others. Figure 19 shows how *year* seems to be the most important feature for the model. This can be explained by the trend in the number of projects and the success rate over the years, previously analyzed in the data exploration paragraph. The fame of the platform after 2014 obviously increased in step with the number of campaigns launched, but many of the latter still did not find enough donors to complete the project. Therefore for the years between 2014 and 2017 there is a decidedly higher probability of failing compared to previous years. The high importance of *time_interval* is justified by the correlation between the success rate and duration of a project; as explained before, a longer time interval can decrease the probability of success (Figure 3).

A more specific categorization seems to bring more information to the model: in the feature importance of the second GBT model, the *category* has an higher values than *main_category*.

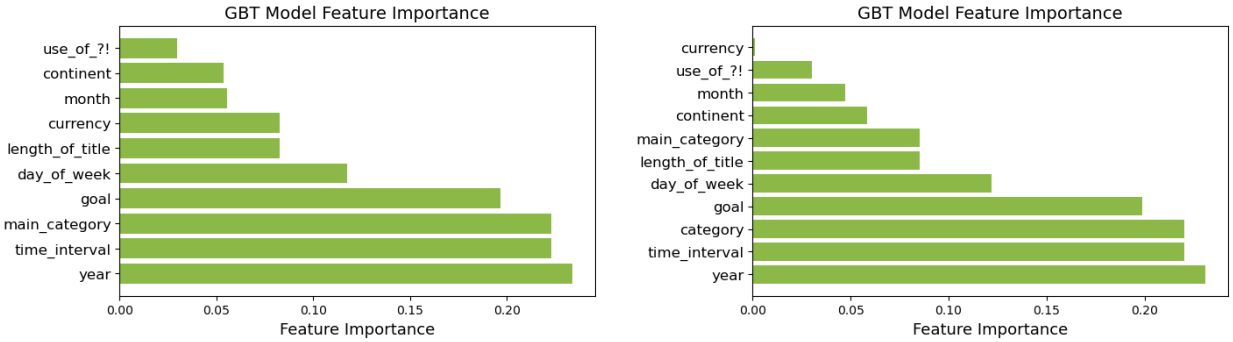


Figure 19: Feature importance as revealed by GBT Classifier

5 Conclusions

Our analysis tells an insightful story about crowdfunding projects from 2011 to 2018. From the data exploration we can conclude that annual and seasonal trends reveal significant shifts in the ratio between fundraisers and backers. Some categories outperform, particularly the niche ones like dance, theater and comics. From the analysis of the title, we can infer that the title words and punctuation do not influence the quantity of people that invest in the project.

The most important variables, and the ones that seems to represent an important impact in the state of a project are goal and the duration set up by the fundraisers. As we could see in the previous sections, a short-medium interval seems to be the best choice, as well as moderate goal values.

Finally, due to the poor performance of the trained classification models used to predict the campaign outcomes, we realized that is challenging to obtain high values of accuracy because of the limited availability of project's information.