Is This Kickstarter Project Going to Raise Its Crowdfunding Goal?

Lior Solomon,Eva Marciano, Omer Sarig, Yotam Aharony

1) Data Description

Our dataset (taken from https://webrobots.io/kickstarter-datasets/) included information about 250k past projects that took place in the years 2016-2019.

We downloaded each year's last CSV (2019-12-12, 2018-12-13, 2017-12-15, 2016-12-15) and merged onto several CSVs. The raw features we got from the original dataset were:

Used features:

- Id – every project in Kickstarter has a unique ID#
- Name – name of the project
- Category – json data that represent the main category and the subcategory of the project
- Location – json data that contains the country and the city origin of the project's creator
- Created at – project's creation time (UNIX time)
- Launched at – project's launch time (when fundraising started) (UNIX time)
- Deadline – project's deadline time (when fundraising stopped) (UNIX time)
- Country – creator's country
- Goal – project's goal (in original currency)
- Currency – project's currency (US, GBP etc.)
- Static_usd_rate – static USD/original currency rate at launch time
- Urls – json data with the address of the project site on the Kickstarter website
- Creator – json data with creator's id and name
- Photo – json data contains project's photo url with different resolutions
- Staff_pick – Boolean value indicating whether or not the project had StaffPick badge (see)
- Blurb – a short description of the project
- **State** - was the project successful or not

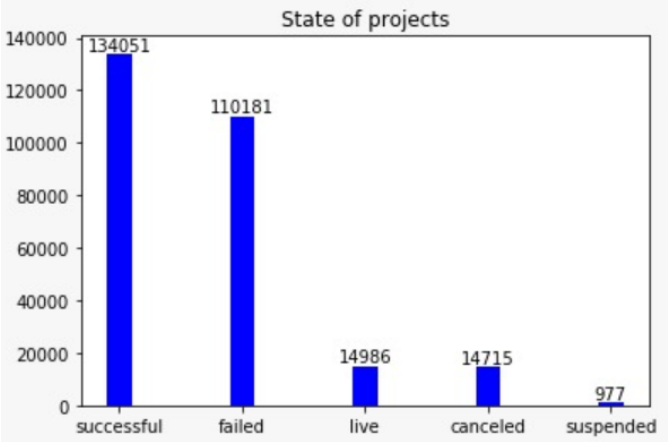
Unused features:

- Pledged – how much money the project pledged
- Slug – project's name in lowercase and without spaces
- Currency symbol – the currency symbol of the project
- Currency trailing code – Boolean variable that was true for USD and false for rest currencies.
- State_change_at – UNIX time of when the project's state was changed
- Backers_count – number of backers

- Usd_pledged – how much money did the project pledged in USD
- Spotlight – was the page made as home for a successfully funded project? (see)

2) Dataset Analysis Summary

- The project's status **successful** rate of our dataset is **55%** (slightly biased. Important: "in the real world", there are only 33% success)



- **Live** - project that still ongoing
- **Canceled** - project that was canceled by the creator before due date
- **Suspended** - project that was canceled by Kickstarter's Trust & Safety team due to violations of Kickstarter rules
- **Successful** - project that was able to raise its crowdfunding goal
- **Failed** - project that **was not** able to raise its crowdfunding goal

Hence, we decided to drop live, canceled and suspended projects – about 15K total projects, which are 5.4% of our dataset.

- Our initial dataset contained about 1000 missing values. As it is negligible to the number of rows in the dataset, we decided to deal with it by drop the rows with missing values.

- Many of the features were obtained as object type. We had to do a conversion to types we could work with. For example, convert the 'Launched at' column from UNIX to a regular DateTime, extracting duration, categories information, convert our goal (to dollar) and more…

- Our data contained projects between 2009 and 2019, based on more than 22 different countries, 15 main categories and 160 sub categories, duration varies from 1-98 days, and goal from several dollars to million.

After diving deep into our dataset, and running a baseline model, we decided to add and extract new features, for example:

- **"Exposure" features -** such as *number of projects launched in the last 7 days from the project launched date (idea from Kaggle kernel)* or *time in hours since last project launched in the same category.* (if a project did not get enough exposure, it may "go down" at the feed and that can affect the project's failure)
- **Country-date information** such as Happiness state and GDP (outsource datasets)
- **"Look-Again" features** - changed the goal column to log scale, extract datetime information such as weekday, month and more.

3) Problem Formulation

Determine whether a new Kickstarter project is going to raise its crowdfunding goal by Machine-Learning Model based on dataset of about 250k past projects.

4) Point of Focus

Nontrivial features extraction

1. Name and Blurb NLP and statistical analysis

   A main parameter that can determine whether a user would want to pledge this project it's how good the project is explained and how attractive its name or blurb. We made some statistical analysis to these two features, such as char count, word count, average word length, number of special chars and more. We also added a feature that defines how many successful words, and how many failed words there are in its name and in its blurb.

   We created a top-100 most common words in successful projects and same for the failed ones. We found out that there are a lot of common words (means – in the both lists).

   A 'successful word' was defined as a word in top-X success but not in the top-Y failed. A 'failed word' was defined similarly. X and Y were tuned to get the best results.

2. Photos

Each project on kickstarter has as main photo that act as the project's "business card. Our dataset has a Photo column, which contains different resolution for the "main photo" of the project. Hence, we extracted and downloaded the best quality image of every project and ran an image quality assessment algorithm on them.

We used a BRISQUE (Blind/*Reference-less* Image Spatial Quality Evaluator) (see). In short, the algorithm calculates the no-reference image quality score for image A using the Blind/Reference-less Image Spatial Quality Evaluator (BRISQUE). BRISQUE compare A to a default model computed from images of natural scenes with similar distortions. A smaller score indicates better perceptual quality. It predicts score by using a SVR model trained on an image database with corresponding DMOS values.

3. Rewards

Kickstarter lets the creator to create a series of rewards to say "thank you" to those who back their project. As the world works, it is more likely that people will invest a project if they are getting something in return. Therefore, we believed that it can significantly affect the success rates of the project.

How can we compare between different rewards of different projects? According to 'Kickstarter' website there are few important aspects while creating rewards:

- What should you offer? What should you not offer?
- How to price
- Offer a range of rewards

We decided to focus on the last two.

After extracting all the rewards from each project's webpage using a crawler, we processed the data and extracted the following features:

- **min** reward price / goal, **max** reward price / goal, **avg** reward price / goal
- num of reward options (i.e. range of rewards)

5) Solution

After data conversions we created a baseline model. Our model contained several basic features: duration, goal, country and main category. We created a SVM linear model and a RandomForest and got these results:

Statistics	
Train Score	55.13%
Test Score	55.87%
Precision	55.78%
Recall	99.56%
fScore	71.50%

Linear SVM

Statistics	
Train Score	73.95%
Test Score	66.88%
Precision	68.33%
Recall	72.09%
fScore	70.16%

RandomForest

We decided to continue with the RandomForest model and extract more features. We examined every feature and we managed to improve our accuracy:

Statistics	
Train Score	86.43%
Test Score	77.39%
Precision	81.08%
Recall	76.82%
fScore	78.89%

After point of focus features, we managed to improve out model in ~3% and get to:

Statistics	
Train Score	89.51%
Test Score	80.25%
Precision	80.81%
Recall	84.06%
fScore	82.40%

After tuning our hyper-parameters, we didn't manage to make a significant improve, we then looked for some more accurate models. Finally, after tuning some hyper-parameters we managed to improve our results in about 4%, by using LightGBM model:

Statistics	
Test Score	83.89%
Precision	85.16%
Recall	85.67%
fScore	85.4229%

Light GBM

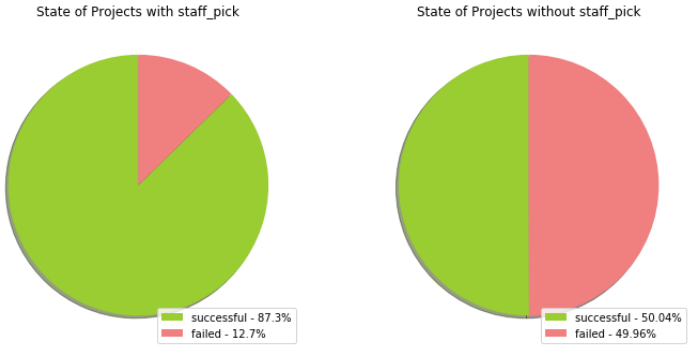
Confusion		Pred	
		No	Yes
Act	No	8927	1998
	Yes	1918	11468

Confusion

6) Findings and Statistical Evaluation

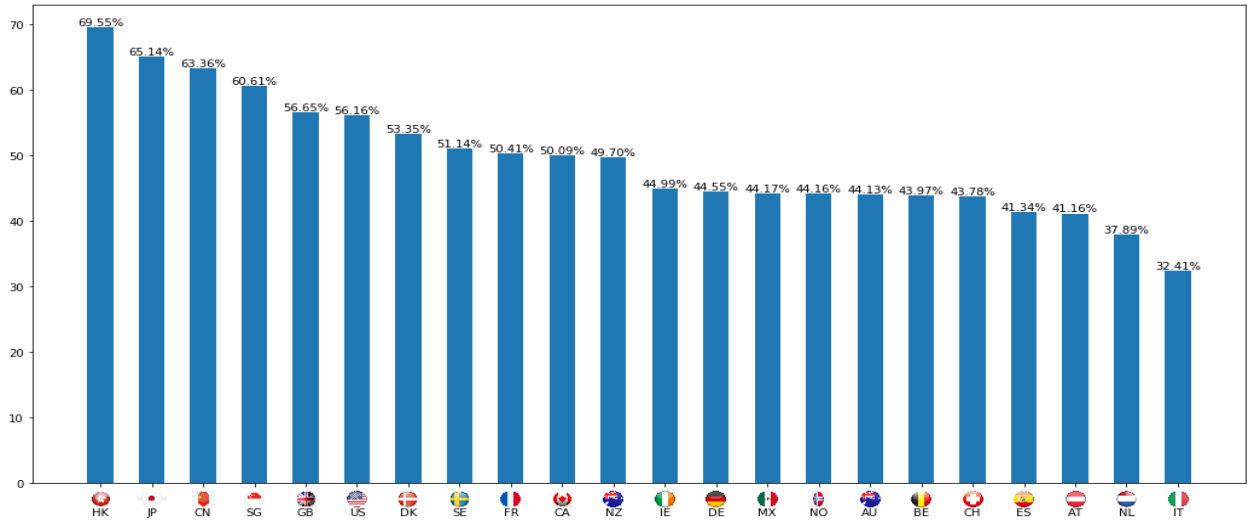
While working on the project, we performed a lot of statistical evaluation that influenced the next step in the project. We shall note a few of them:

- **Staff Pick -** We can see that the staff_pick has a big influence on the success of a project. Indeed 87.3**%** of projects with staff_pick are successful,while only 50% of projects without staff_pick are successful.

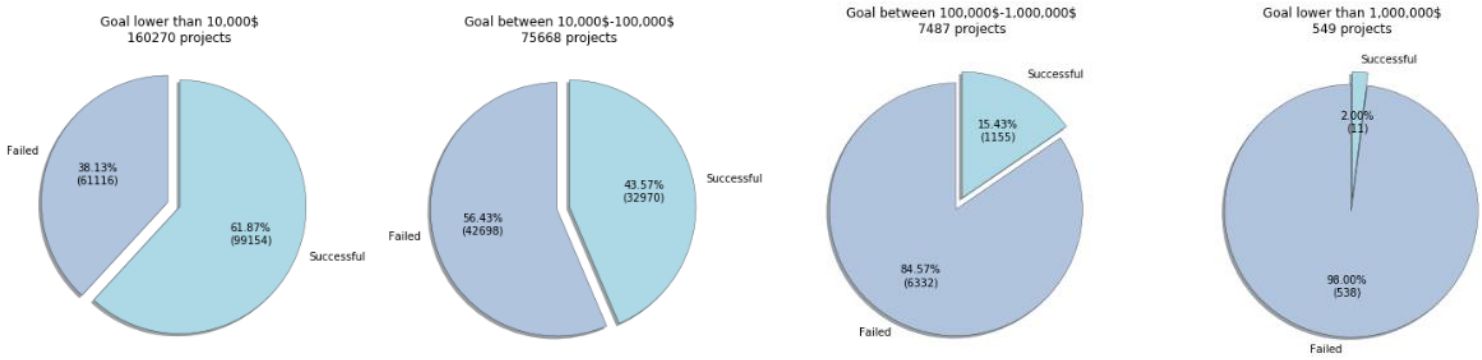


- **Country -** We can see that the country has a good correlation with the success rate as can be seen in the diagram below. them in descending order of success rate. We can see that there may be a large difference in the success rate between different countries (~40% difference between Hong Kong and Italy)

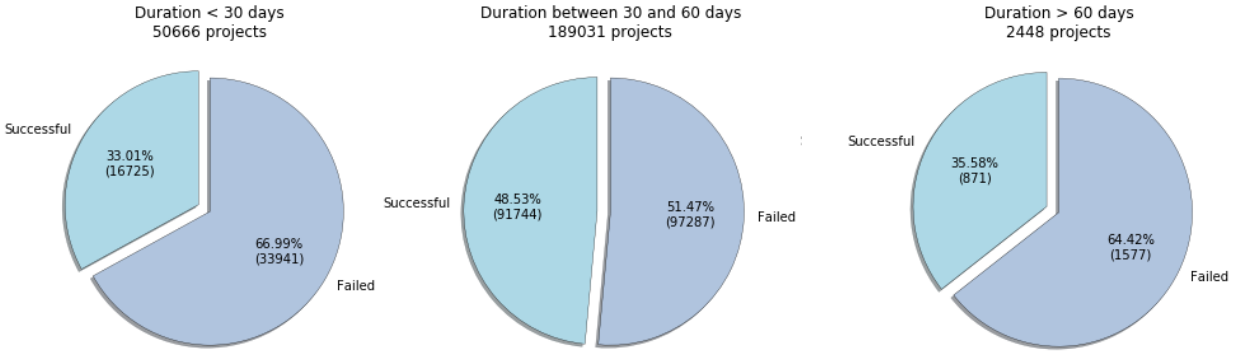
-



- **Goal -** We can notice a correlation between the goal and the success of a project. a project with a lower goal is more likely to succeed.



- **Duration -** there's a strong correlation between the duration of the project and its success. Contrary to what we might think, the longer the duration, the less the project is likely to succeed.



7) Insights and applications

- As can be seen it's a complicated problem, which can have a related solution by boosting and majority of several models (therefore RandomForest and LightGBM gained the best results).
- According to the importance of features on the LightGBM model, we can say that the 5 categories that most affect the success of a project are sub_category, log_usd_goal, max_reward_price, avg_reward_price and count_7_days. Therefore, if a creator wants to create new Kickstarter project, he may focus on those 5 categories.

- We would also want to recommend of a goal of under 10,000$ (if possible), 30-60 days duration, name contains successful words such as documentary, edition and volume, and do not contain failed words such as dream, real, people, app and web.
- We can offer our project to Kickstarter to improve the success rate of each project, by running our model on it pre-launching in order to get a good estimate if it's going to succeed. Furthermore, we could have improve our project to get hints and tips of how making this specific project to a success.
- Our research can be expanded to a regression problem and predict how much money will be pledged for each project.

8) Related Work

The idea for our project was taken from Kaggle website, therefore previous work was done. Most of the 'Kernels' on Kaggle has managed to reach about 70% accuracy (specific kernels attached below). Unlike Kaggle's dataset, our dataset was taken from the WebRobots website which contains more information then the Kaggle dataset.

Our task was to get a better performance from the Kernels on Kaggle and we chose to do so by adding some non-trivial features extraction as explained in 'Point of focus' section.

9) Citations

- Kickstarter Kaggle
- Dataset source
- Ideas for rewards analyzing
- Kernel from Kaggle website on data cleaning and normalize data.
- Kernel from Kaggle website on the feathers: Number of projects in the last 7 days from project launch date and time in hours since last project in the same category.
- Photos BRISQUE algorithm.
- Happiness Kaggle source.
- GDP Kaggle source.
- Python libraries:
    - NLTK
    - Country converter
    - Word cloud

10) Working with Friends:

While working together with our colleague team, we decided to base our collaboral work on their project, and sharing our different features (by external CSVs or by sharing the code). We “extracted” our features:

- Rewards Features (Min/Goal, Max/Goal, AVG/Goal, #number_of_rewards)
- Photo Score
- Count7Days
- CategoryCount7Days
- Log Goal

By that, We managed to improve our combined accuracy by 6% from 84% to about 90%, with 0.81 Recall and 0.9 precision (!!!).