Predicting Success of Kickstarter Campaigns

Mark Pothier

Abstract

This project studies a dataset of nearly 100,000 Kickstarter campaigns with details including funding goals, duration, categories, and location. The primary question: **how well can we predict if a Kickstarter campaign will be successful from the outset?**

Using tools from the *numpy*, *pandas*, and *scikit-learn* Python libraries, we will study which features play larger roles in the success of a campaign, and ultimately we will build two machine learning classifiers in seeking to maximize our prediction accuracy.

At the conclusion, we learn that a few features stand out as the most important to determining campaign success: **fundraising goal**, creative **category**, label as a **Staff Pick**, and **time spent creating** campaign.

Motivation

Kickstarter is an online funding platform for people trying to get their personal projects off the ground. Projects fall into a variety of categories and usually offer a reward system for donors that reflects their pledge amount.

The twist is that funding is **all or nothing**. If a project falls short of its funding goal at the pre-selected deadline, the project fails and the creator receives **no funding** (all pledges are returned to their respective donors). Therefore both campaign creators and donors alike have a strong incentive to make sure that their projects have the best chance of success possible!

Datasets

Two datasets were used in this study, and ultimately merged to create a more feature-rich dataframe. Both datasets were found on Kaggle:

- Dataset 1: https://www.kaggle.com/wood2174/mapkickstarter (~99,000 entries)
- Dataset 2: https://www.kaggle.com/kemical/kickstarter-projects (300,000+ entries)

The two datasets were merged on the campaign ‘ID’ and have a fair amount of overlap; consolidated features include main category, sub-category, status (successful or failed), city, state, fundraising goal ($), month launched, campaign duration, and campaign prep-time.

The data was filtered for U.S. projects only, in order to ensure that the fundraising goal values were all normalized for currency. While studying Kickstarter campaigns around the country would be an interesting follow-up study, it should be noted that a vast majority of Kickstarter campaigns originate in the U.S., and therefore our sampling is appropriately representative of Kickstarter overall.

Data Preparation and Cleaning

The datasets were reasonably well-organized .csv files; however some clean-up was required to eliminate unnecessary/duplicate feature, re-name features, and convert boolean or string features to numbers for the purposes of classification.

Additional cleaning and preparation was required in order to merge the two datasets, ensuring they aligned on the intersection of their unique IDs, and removing entries with null values after merging.

Research Questions

1. If I'm looking for campaigns to support, how can I determine which are most likely to succeed?

    -Are some categories more successful than others?
    -How accurately can I predict if a campaign will be successful?

2. If I'm starting a campaign, which factors should I focus on to increase its chance of success?

    -Does it matter where I'm located?
    -Does it matter in which month my campaign is launched?
    -Does the length of the campaign make a difference?

Methods

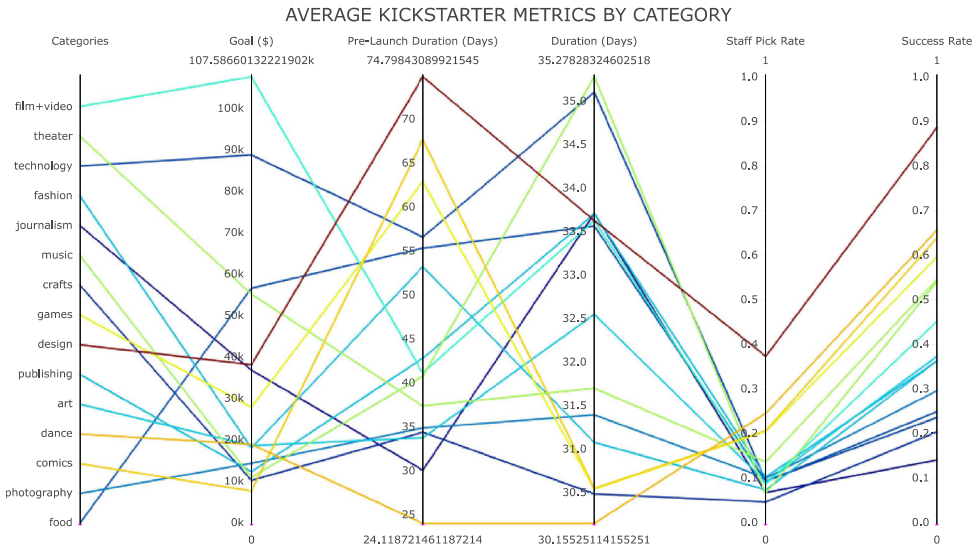
As previously mentioned, this study used tools from the *numpy*, *pandas*, and *scikit-learn* Python libraries.

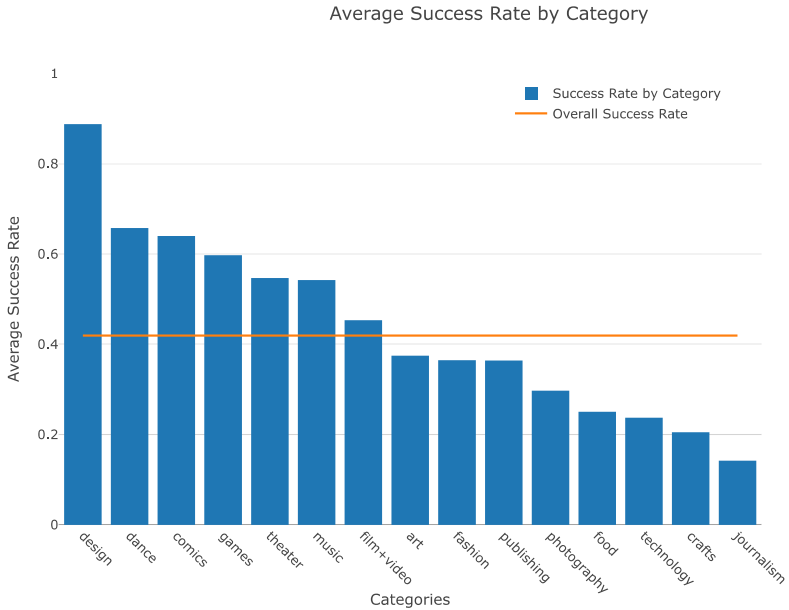
With basic *numpy* and *pandas* tools, I tested out hypotheses by feature (e.g. grouping entries by categories, or by staff pick) to understand general trends.

With *scikit-learn*, I built both Decision Tree and Random Forest classifiers to calculate prediction accuracy and learn ways in which both models can be tuned to enhance their accuracy. Both classifiers were run with GridSearchCV and simple cross-validation to increase our confidence in the prediction accuracy.

All visualizations were created with the *Plotly* library.

Findings





The data demonstrated that campaigns vary widely across a number of features (including success rate) when grouped by category.

Overall success rate averaged out to **42%**; however this varied widely among categories, from ‘design’ at **89%** to ‘journalism’ at just **14%**.

Findings

The designation of a campaign as a “Staff Pick” - which effectively means it is promoted by Kickstarter on their website - yields a decisive advantage (**84% vs 37%** success rate).

	staff_pick	id	goal	status	Length_of_kick	Days_spent_making_campign	City_Pop	Cat-Nums	City-Nums
0	0	1.075390e+09	36891.891039	0.372683	33.949108	42.895471	1.107842e+06	7.840833	1205.229543
1	1	1.080529e+09	21865.249617	0.835791	32.747198	63.601864	1.763350e+06	7.433054	1135.564940

Notice that two features offer possible insights into how campaigns are chosen as Staff Picks: **fundraising goal** and **length of campaign**. The data appears to indicate that more modest fundraising goals, combined with more time setting up the campaign, increases a project’s chance of being a Staff Pick.

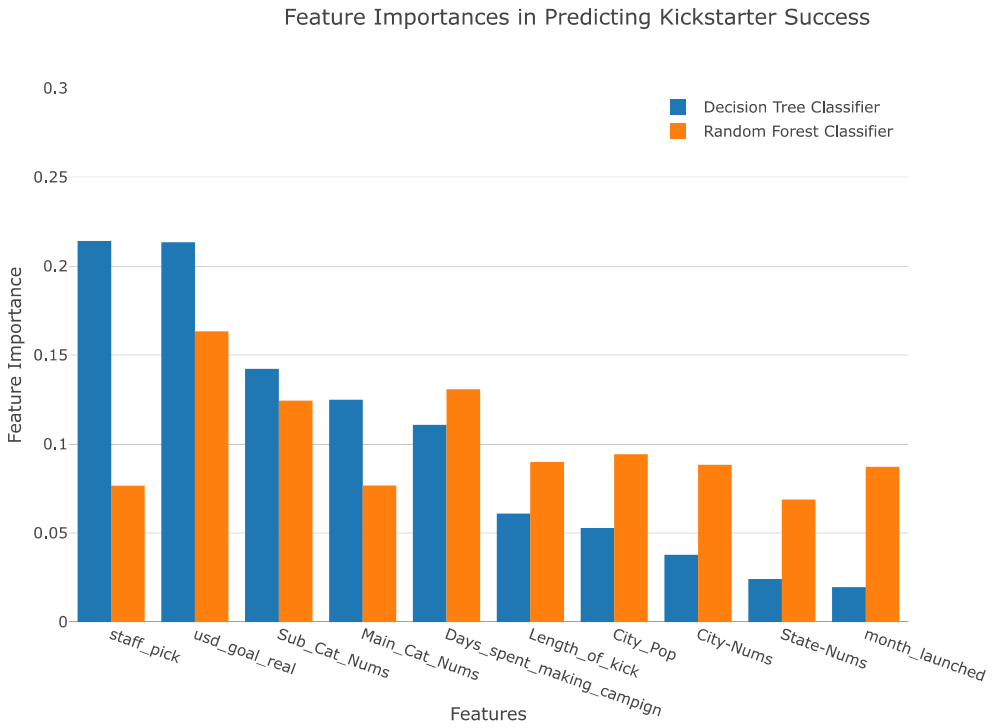
Findings

Both a Decision Tree and Random Forest classifier were used to calculate accuracy for predicting the success of a campaign, with the following results:

- Decision Tree Classifier: **72%**
- Random Forest Classifier: **74%**

From each model, we learned the relative importance of each feature from our dataset (right). Among the most important features:

- Staff Pick
- Fundraising goal
- Sub-category
- Time spent preparing campaign



Limitations

Note that the dataset was limited by common IDs between the two initial datasets, and was further limited to U.S. campaigns (due to missing currency exchange data).

There are additional data features that would give a more complete picture of campaigns and that I would like to include in the future to move this study to the next level:

- Suggested pledge amounts
- Number of different suggestions
- Pledge suggestion as percent of goal
- Rewards offered / reward thresholds
- Campaign page contents (videos, photos, charts, graphics, description, etc.)

This data may be readily available in another dataset that I haven't yet discovered; otherwise, it can likely be web-scraped from an archive of concluded campaigns from Kickstarter.

Conclusions

1. *If I'm looking for campaigns to support, how can I determine which are most likely to succeed?*

   One should first note that a campaign's creative category plays a large role in its chance of successful funding. Additionally, searching among the Staff Picks will put you in a pool more likely to succeed right off that bat.

   In terms of prediction, our models can predict campaign success with up to **74%** accuracy.

2. *If I'm starting a campaign, which factors should I focus on to increase its chance of success?*

   Location and launch date don't appear to play a large role, but if a creator wants to give their campaign's chance of success a huge boost, they should aim to earn their project a Staff Pick (and the resulting promotion). In order to do so, a more modest fundraising goal and well-planned campaign will increase those odds.

Acknowledgements

All data was collected from datasets publicly available on Kaggle.

No outside feedback was received at the time of completing this study.

References

Additional insights into Kickstarter trends were gleaned from Kickstarter's blog:

- https://www.kickstarter.com/blog/trends-in-pricing-and-duration

Various technical guidance was received via Google, Stack Overflow, Scikit-Learn documentation, Plotly documentation, and Git Hub forums.