The Key to a Successful Kickstarter:
Predicting the Amount of Money a Kickstarter Campaign Will Raise

Abstract: Crowdfunding is the act of raising monetary contributions to fund a project or a venture. In this project, we explore the various factors that influence the amount raised by crowdfunding campaigns on Kickstarter, arguably the biggest crowdfunding platform. We obtained the data, collected in 2012, from BigML, a machine learning repository. We use the best subset regression technique to find the optimal factors that predict the desired response and build a multivariable linear regression model using those predictors. The initial model created predicts the amount raised by Kickstarter campaigns in 2012 with an R-squared of 85% and our final model has an R-squared value of 85.1%. We then conduct an extra sum of squares test from which concluded that presence of the 'has_video' variable significantly improved our model.

Background and Significance

Crowdfunding, often performed online, is one of the easiest and fastest ways through which individuals or groups raise monetary contributions for their ventures or projects. Many developers, musicians and property owners have turned to crowdfunding for their projects, since banks or other financial institutions have turned them down. As online crowdfunding has become a dominant medium for financing projects, we believe it is important that we understand the factors that a necessary to drive a successful crowdfunding project on Kickstarter - the chosen medium for our research. Research cited in an article by AdWeek[1] shows Kickstarter as the best platform for crowdfunding thus it was chosen as our medium of interest. By examining numerous Kickstarter campaigns in the United States, we seek to find the factors that influence the amount of money raised by these campaigns in the year 2012, in order to better tailor future campaigns for successful outcomes.

Methods

Data Collection And Trimming

For our study, we obtained the dataset from BigML[2], a machine learning data repository. Further details on the collection of the data can be found on the website. Initially, we had 49 variables in our dataset. As part of cleaning the data, we chose to exclude 26 variables that we thought were not necessary for the purpose of our study (this is discussed in details in the Appendix). We also removed all incomplete entries in the dataset to ensure that we were working with a completed entries without having to make unfounded assumptions.

Variable Selection

We conducted a best-subset regression using the “leaps” package in RStudio; we found the best combinations of variables that are good predictors for the response we seek. After viewing some residual plots, we applied logarithmic transformations on some variables (eg. “pledged”) in order to normalize the distribution of the data. From the results of our best subset regression, we create various linear models and compare them to find the best fitting model with the least variables. We attempted several interaction terms as well, but were unable to any significant ones. We go on to use four variables from the many since the resulting model with these variables gave a reasonable R-squared value of 0.85 and there was an insignificant difference between that R-squared value and that of others with even more variables.

Our initial model:

log(pledged) ~ log(goal) + log(backers count + 1) + log(comments count + 1) + deadline month

[1]Source: Kickstarter vs Indiegogo: Which is Best for your Crowdfunding Campaign?
http://www.adweek.com/socialtimes/kickstarter-vs-indiegogo-best-crowdfunding-campaign-infographic/205069

[2]Source: Kickback Machine
https://bigml.com/user/jdonaldson/gallery/model/50c5446c035d074bc1000110

Analytic Methods

The outcome of interest for our study is the amount of money raised by campaigns. Before modelling, we suspected that campaigns with descriptive videos may may influence the amount of money these campaigns raised thus we conducted an extra sum of squares test to validate or reject our assumption. From our test, we found that the full model has an F-statistic of 34.3 under 1 degree of freedom and a p-value of 4.8e-09. Thus, we reject the null hypothesis that “has_video” is not an important predictor and conclude that the “has_video” variable is significant and should be included as a predictor in our model. The new R-squared of our model including the has_video variable is 85.1%. Our full model is shown as follows:
log(pledged) ~ log(goal) + log(backers count + 1) + log(comments count + 1) + deadline month + has video

Table 3: Summary statistics of variables for full model

	Estimate	**Pr(<|t|)**
Intercept	2.10368	<2e-16
log(goal)	0.15956	<2e-16
log(backers count) + 1	1.30249	<2e-16
log(comments count) + 1	-0.18800	<2e-16
deadline month	-0.06368	<2e-16
has_video	0.13881	4.8e-09

Results

Below are scatterplots of the variables against the response and colored by has_video variable

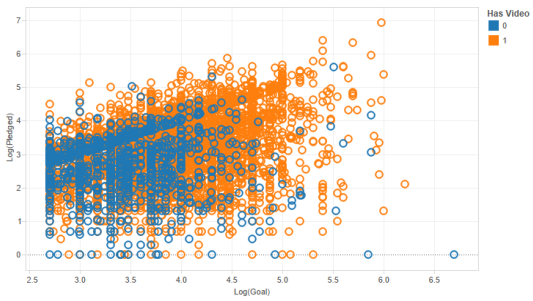

Figure 1: log(pledged) vs log(goal)

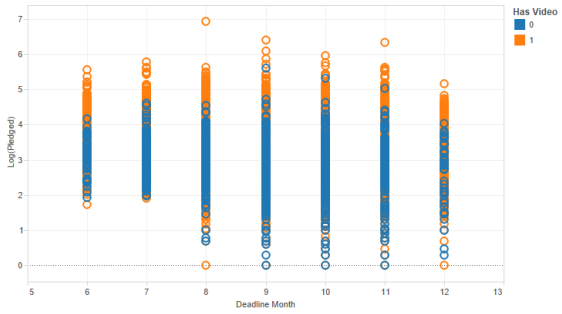

Figure 2: log(pledged) vs deadline.date_month

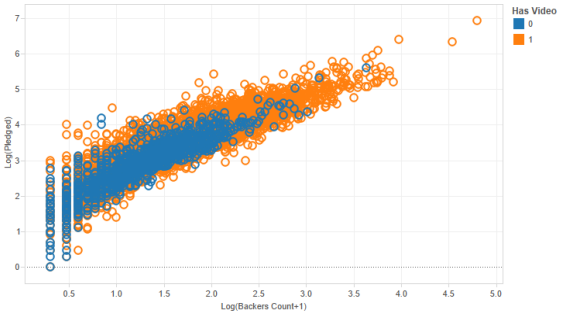

Figure 3: log(pledged) vs log(backers_count)

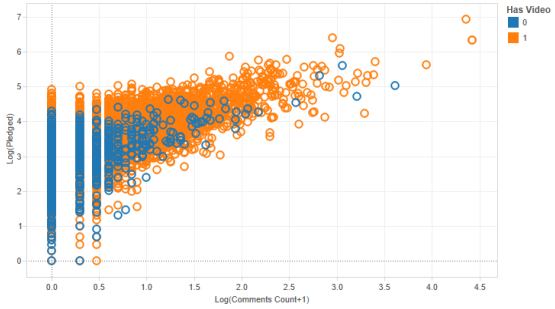

Figure 4: log(pledged) vs log(comment_count)

We seek to answer the question, “What factors influence the amount raised by a Kickstarter campaign?”. Considering that our response is numeric, we choose to use a multivariate linear regression model. After selecting our variables and confirming our hypothesis about the relevance of videos in Kickstarter campaign, we built our model. With an R-squared value of 85.1%, we are confident in the result of our model’s prediction and also the fit of the data to our model.

Discussion and Conclusion

Our objective was to find factors that influence the amount raised by a Kickstarter campaign in 2012. In particular, the use of descriptive videos in Kickstarter campaigns Using best subset regression, we are able to create a strong model with only five terms. Since our dataset was collected over a short span within a specific year, we are not able to build a generalized model. Thus, it is important to note that our model cannot be generalized to make future predictions.

Our analysis is important in that one learns about how the found predictive factors influenced the amount pledged to Kickstarter campaigns in 2012. It will be interesting to see the modelling of trends in a wide-span dataset (i.e one that covers a wide range of years). Thus, for future research, we hope to obtain data spanning a longer range of time to be able to model any change in trends. Nonetheless, we believe there are numerous possibilities for questions and different explorations that could be done with this dataset.

References

Mollick, Ethan R., The Dynamics of Crowdfunding: An Exploratory Study (June 26, 2013). Journal of Business Venturing, Volume 29, Issue 1, January 2014, Pages 1–16. Available at SSRN: http://ssrn.com/abstract=2088298 or http://dx.doi.org/10.2139/ssrn.2088298

Robertson, Emily Nicole and Wooster, Rossitza B., Crowdfunding as a Social Movement: The Determinants of Success in Kickstarter Campaigns (July 15, 2015). Available at SSRN: http://ssrn.com/abstract=2631320 or http://dx.doi.org/10.2139/ssrn.2631320

Appendix

Description of Variables Present in the Dataset

We have the following variables in our dataset:

- pid : process identifier
- scraped: a boolean, indicating that the data for a given row was scraped.
- scaped_datetime: time the data was scraped
- thumbnail_url: url for image as thumbnail
- short_url: shortened url
- url : The url for the crowdfunded project on Kickstarter
- name: Name of the campaign
- percent_raised: Portion of the goal raised
- goal: Target amount to be reached
- success: boolean, determines whether a campaign was successful or not
- pledged: amount pledged to the campaign
- parent_category_string: category of the campaign
- reward_count: the number of reward levels offered to backers
- update_count: The number of times there has been an update about the campaign
- location : location of the campaign, city and state
- backers_count: number of backers of the campaign
- category_string: narrowed category of the campaign by type
- duration: how long the campaign run
- comments_count: the number of comments
- firstseenin: the category in which the campaign was first seen in (i.e Recently launched, ending soon, none)
- avg._pledge_per_backer: the average amount pledged by backers
- has_videos: whether the campaign has a video or not
- facebook_connected: boolean, whether the campaign has a facebook link or not
- facebook_friends: number of facebook friends
- reward_level_list: List of values of rewards offered for backers
- twitter_followers: the number of twitter followers
- currency: the currency in which the money is collected
- pledged_in_USD: amount pledged in USD
- deadline_date.day-of-week: day of the week of the deadline deadline
- deadline_date.year: year of the deadline
- deadline_date.month: month of the deadline
- deadline_date.day-of-month: day of the month of the deadline
- launched_date.year: Year of the launch date
- launched_date.month: month of the launch date
- launched_date.day-of-month: day of the month of the launch date
- launched_date.day-of-week: day of the week of the launch date
- duration(months): duration of the campaign
- ks_id: kickstarter id
- kickstarter_fee: amount of money taken out when by Kickstarter when the project is successfully funded

Variables Removed in Cleaning:
In trimming the dataset for analysis, we removed the following variables which were either not of interest for our project or not formatted in a way that we were able to utilize them:

- deadline_date : deadline date for the crowdfunded project
- deadline : the deadline date and time for the crowdfunded project
- pid : process identifier
- scraped: a boolean, indicating that the data for a given row was scraped.
- scaped_datetime: time the data was scraped
- thumbnail_url: url for image as thumbnail
- short_url: shortened url
- first_seen: time the campaign was first seen
- full_description: full text description of the campaign
- short_blurb: a brief description of the campaign.
- launched: the launch date and time for the crowdfunded project
- facebook_link: the url to the facebook page
- twitter_screenname: twitter account name
- project_state: the status of completion of the project (i.e cancelled, failed, successful)
- launch_date: the date the project was launched.
- twitter_url: the url to the twitter account of the project
- ks_id: kickstarter id
- kickstarter_fee: amount of money taken out when by Kickstarter when the project is successfully funded
- facebook_connected: boolean, indicating whether the project is linked to facebook
- duration: unknown calculation of the length of the project