How to succeed in crowdfunding: Setting a perfect goal

Thanh Tran
Department of Computer Science
Utah State University
thanh.tran@aggiemail.usu.edu

Hongkyu Choi
Department of Computer Science
Utah State University
hongkyu.choi@aggiemail.usu.edu

ABSTRACT

1. INTRODUCTION

2. RELATED WORKS

3. ANALYSIS

4. FEATURES

In this section, we propose some features which are useful to develop two predictors of how much funding and the number of backers. We groups all our proposed features in different traits:

4.1 Project-based features

Given a Kickstarter project's page, we extracted all following features:

- number of project's comments,
- number of updates
- number of rewards,
- number of images,
- number of videos,
- number of faqs,
- the goal of the project
- category of the project
- duration of fund raising of the project
- smog score of all rewards: we extracted all rewards' description of the Kickstarter project then grouped them into one document. Next, we computed the smog score of the document and considered it as smog score of all rewards' description.
- number of reward sentences,
- number of main sentences,
- smog score of project's description: we extracted the project's description and computed its smog score.
- smog score of biography description

SMOG score of a document show how well the document is written. The higher the SMOG score, the better the document was written. The formula to compute SMOG score is as following:

$$grade = 1.0430\sqrt{\#polysyllables * \frac{30}{\#sentences}} + 3.1291$$

4.2 Creator-based features

We proposed some features related to Kickstarter creators as following:

- successful rate of all previous backed projects: given a Kickstarter project, we collected all projects that the creator backed and calculate the average successful rates of these backed projects. Our intuition behind it is that if the creator backed many successful projects, he may have an idea of how to get more fund and attract more backers.
- previous successful rate of creator's project: we first collected all projects that the creator launched. Then we calculate average successful rate of such projects. The intuition is that if the creator successfully raised fund in many projects, he may have many experience in raising more fund or getting interest from backers.
- number of created projects: as mentioned in [1], the more projects that creator launched, the more successful he is in raising fund.

4.3 Social network - based features

Social promotion is an effective way to promote Kickstarter project and attract more funding. Hence, we use some features related to social communication of the creator as following:

- number of Facebook friends
- connected to Twitter? (binary feature)
- connected to Facebook? (binary feature)
- connected to Youtube? (binary feature)
- how many websites that the creator has?

WOODSTOCK '97 El Paso, Texas USA
© 2016 ACM. ISBN 123-4567-24-567/08/06...$15.00
DOI: 10.475/123_4

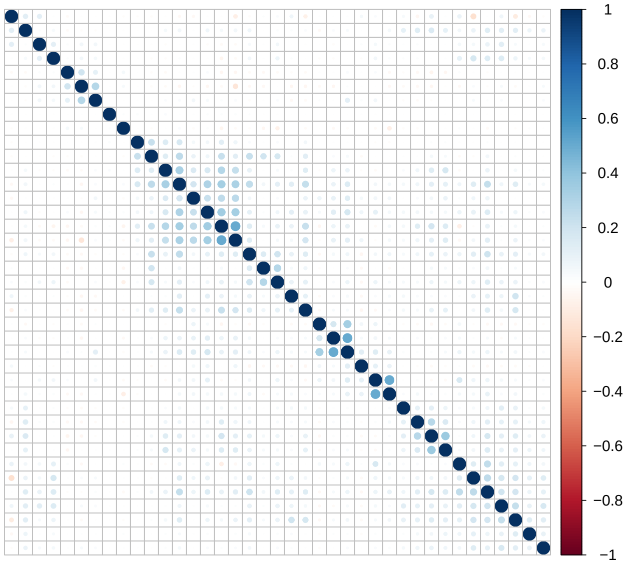


Figure 1: Correlation matrix of all features after removing highly correlated features

4.4 Feature Selection

In this section, we present our approach to remove non-significant features. We first remove all highly correlated features by setting cutoff value of 0.7. That is, if a feature has correlation score of greater than or equal to 0.7 with another one, we remove such feature. After removing all the highly correlated features, we obtained the correlation matrix as shown in Figure 1.

Among all the rest of uncorrelated features, we would like to keep only significant features. In order to do so, we sequentially choose p features (p = 1,..,n) and calculate the stepAIC value as following:

$$stepAIC = nln(SSE_p) - nln(n) + 2p$$

where n is the total number of features, SSE_p is the sum of square error (SSE) that we got by building the model with p features. We will choose the subset of all n features which minimize stepAIC score.

5. OUR PROPOSED NON-LINEAR REGRESSION MODEL

Fitting a linear model for our problems of predicting the number of backers and the amount of funding for a given Kickstarter project may not be always helpful since the correlation between all features and response could not be linearity. Hence, we proposed a non-linear model as following:

$$Y^{\lambda} = \beta_0 + \beta^T X + \epsilon \tag{1}$$

Here, ϵ is the random error and follows normal distribution $N(0, \sigma^2)$. The power of this model is that with different value of λ, we have different curvature function. For instance, when $\lambda = 0.5$, we have the function:

$$\widehat{Y}^{1/2} = \beta_0 + \beta^T X = X\beta \tag{2}$$

The Equation (2) can be rewritten as following:

$$\widehat{Y} = (\beta_0 + \beta^T X)^2 = (X\beta)^2 \tag{3}$$

Now, the Equation (3) is a quadratic function, not a linear model. Similarly, when $\lambda = 1/3$, we have a cubic function that get through all points.

The next problem is how we can find out the best value of λ for the model? In order to seek the answer for that question, we first find out the likelihood function of our proposed model. Given features' value and the real value Y, our proposed model generated predicted value $\widehat{Y}$. Our goal is to minimize the optimized function of error:

$$\min error^2 = \sum_{i=1}^{n} (Y_i^{\lambda} - \widehat{Y}_i)^2$$

We assume that the error follows normal distribution N(0, σ^2). The probability density of $error_i$ of ith observation is given as following:

$$\begin{aligned} f(error|\mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{error_i^2}{\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i^{\lambda} - \widehat{Y}_i)^2}{\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i^{\lambda} - X\beta)^2}{\sigma^2}} \end{aligned} \tag{4}$$

Given Equation (4), the loglikelihood function of error is:

$$\begin{aligned} log(L) = -\frac{n}{2}kig(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[Y_i^{\lambda} - X\beta]^2 \\ +(\lambda - 1)\sum_{i=1}^{n} log(Y_i) \end{aligned} \tag{5}$$

Here, σ can be estimated by MSE obtained in the regression model. And the problem of finding the best value of λ will become the problem of maximizing loglikelihood function given in Equation (5).

6. EXPERIMENTS AND RESULT

In this section, we described all our experimental settings which are used in following sections for predicting how many backers will back for a certain Kickstarter project and how much funding that the project can receive.

6.1 Experiment Setting

Dataset: With all 151,608 projects, we see projects with extremely high amount of funding or number of backers as outliers and remove them. We also consider projects with the goal less than \$100 as noisy projects and should remove them. As a result, from 151,608 projects as origin, we filtered out 36,731 outliers and the rest contains only 114,877 projects.

Measure: In all below experiments, we evaluate the result based on root mean square error ($RMSE$). $RMSE$ is a common measure to evaluate the average difference between the predicted values and real values. The formula to compute RMSE is given as following:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n}}$$

where Y_i is the real value of response at observation i and $\widehat{Y}_i$ is the estimated value of response at observation i.

Features Normalization: Since different features have different scale, the features with larger scale may have lower coefficient then ones with smaller scale and may cause larger error. In order to avoid the problem, we normalized all features into [0,1] scale by applying softmax normalization. The formula of softmax normalization is given as following:

$$x_{ij}\prime = \frac{1}{1 + e^{-\frac{x_{ij}-\mu_i}{\sigma_i}}}$$

where x_ij is the value of *ith* feature at observation *j*, μ_i and σ_i are the mean and the standard deviation of all the values of *ith* feature, respectively.

Predicting how many backers will back for a Kickstarter project and how much funding the project can receive: In this experiment, we used our proposed features and build the models to predict how many backers will back for a Kickstarter project and how much funding the project can receive based on 3 approaches: linear regression, extreme gradient boosting for linear regression and our regression model.

6.2 Experiment Result

For our proposed regression model, we first find out the best value of λ by maximize the loglikelihood function given above section

Figure 2 shows the change of loglikelihood when varying the λ value. For building the model to predict number of backers, we observed in Figure 2(a) that at $\lambda - 0.2$, the loglikelihood function achieve the highest value. With regard to the model of predicting the amount of funding, it is shown in Figure 2(b) that we obtained the maximum value of loglikelihood when $\lambda = 0.18$. Hence, we set $\lambda = 0.2$ and 0.18 in the models of predicting number of backers and the amount of funding, respectively.

As a result, we fit the following model for predicting number of backers:

$$Y_{backers}^{0.2} = \beta_0 + \beta^T X + \epsilon \tag{6}$$

And the model for predicting the amount of funding need to be fitted is as following:

$$Y_{funding}^{0.18} = \beta_0 + \beta^T X + \epsilon \tag{7}$$

where ϵ is the random error and $\epsilon \sim N(0, \sigma^2)$

Predicting how many backers will back for a Kickstarter project: we fit our proposed model given in Equation (6). Then we compare the result with linear regression model and extreme gradient boosting for linear model. Figure 3(a) shows RMSE of 3 models for predicting how many backers will back for a Kickstarter project. We note that linear regression model obtained lowest result with RSME of 66.9. Boosting linear regression with Xgboost gained higher result with RSME of 28.7. However, comparing to these two models, our approach obtained the best result with lowest RSMSE of 23.7.

Predicting the amount of funding a Kickstarter project can receive: we fit our proposed model given in Equation (7) to predict how much funding a Kickstarter project can be funded. We also compared our result with the results we got from linear boosting of Xgboost and linear regression. Figure 3(b) shows RMSE result of 3 models. We notice that our model outperformed linear regression and Xgboost for linear model with lowest RMSE of 1,659. This value is the average error of predicting the amount of funding for a Kickstarter project.

7. CONCLUSION AND FUTURE WORK

In this paper, we addressed two problems of predicting how many backers will back for a Kickstarter project and how much funding that the project can receive. For the sake of seeking the answers for these problems, we proposed a non-linear model which made use of linear regression model after doing transformation on response *Y*. We compare our model with linear regression model and extreme boosting model for linear regression (Xgboost) which is a state-of-the art linear model. In our experiment, we have shown that our model is outperformed the rest with lowest RMSE in both two problems: RMSE of 23.7 in predicting how many backers will back for the project and RMSE of 1659 in predicting how much funding the project can receive. Our work provides new insight into crowdfunding campaigns, that help creators set a perfect goal for their projects and increase their successful rate in raising fund.

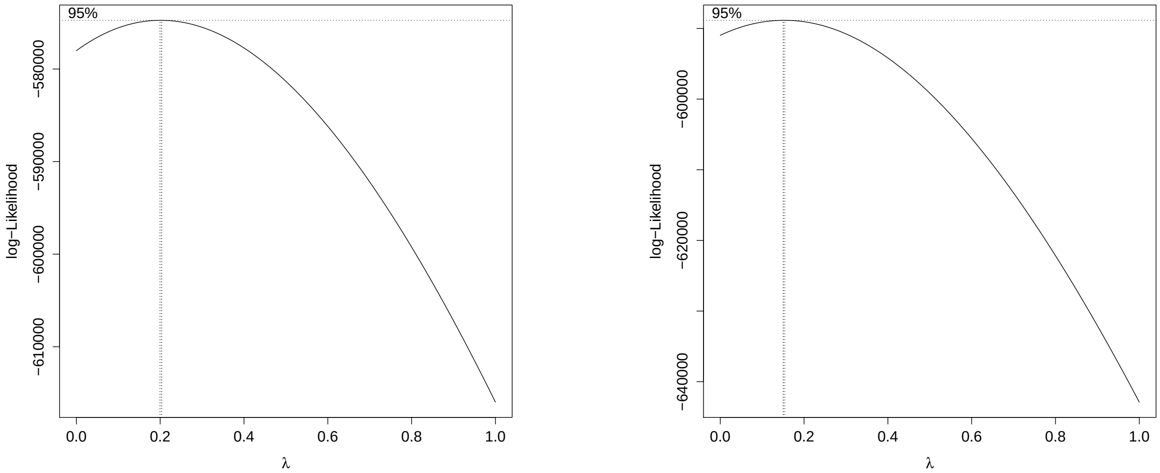
Our work is still limited into Kickstarter platform only. In the future, we will collect all crowdfunding projects in different platforms and suggest more general features. We also improve the model so that RMSE value can be smaller.

Keywords

ACM proceedings; LaTeX; text tagging

8. REFERENCES

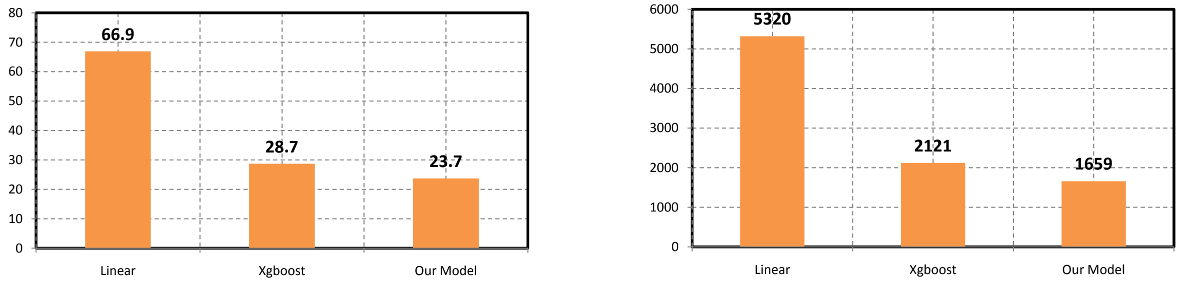
[1] J. Chung and K. Lee. A long-term study of a crowdfunding platform: Predicting project success and fundraising amount. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 211–220. ACM, 2015.



(a) Value of λ in regression model for predicting number of backers

(b) Value of λ in regression model for predicting amount of funding

Figure 2: Value of λ in 2 models: predicting number of backers and the amount of funding



(a) Predicting how many backers will back for a Kickstarter project

(b) predicting how much funding a Kickstarter project can receive

Figure 3: RMSE of different models for (a) predicting how many backers will back for a Kickstarter project, and (b) predicting how much funding will a Kickstarter project receive