Setting a perfect goal to succeed in Kickstarter: A non-linear regression model

Thanh Tran
Department of Computer Science
Utah State University
Logan, UT 84341
thanh.tran@aggiemail.usu.edu

Hongkyu Choi
Utah State University
Utah State University
Logan, UT 84341
hongkyu.choi@aggiemail.usu.edu

ABSTRACT

To find the answer of how to succeed in crowdfunding, we find many previous works which are mainly focused on to recommend appropriate investors to a project or investigate more influential features on its success. However, in this paper, we would like to suggest a way to set a perfect goal with introducing a regression model. Establishing a goal appropriate is essential for project to success eventually as well as predicting the number of backers a certain Kickstarter project.

Our proposed model show better performance than existing algorithm(Simple linear regression and XgBoost). It returns RMSE(Root Mean Square Error) 23.7 and 1,659, prediction for the number of backer and funding respectively.

Keywords

Kickstarter; crowdfunding

1. INTRODUCTION

Why is this an interesting question to ask and why would we care about the answer to this question or a solution to set a perfect goal? The Kickstarter policy is one or nothing. It means that if the amount of final pledged money is equal or higher than the goal, the fundraising campaign is successful. Otherwise, the fundraising campaign is fail and the creator receives nothing. If the creator overestimate the project goal, they will be fail, but if the creator set the project goal too low, they may not get attraction from community and gain less amount of money. Hence, predicting the amount of money or the number of backers are very important for Kickstarter creators to raise fund.

To our knowledge, There is only one research work try to predict how much fund a Kickstarter project can receive. Particularly, researchers at [4] convert the problem of predicting the amount of final pledged money into classification problem by dividing such amount of final pledged money into different range. So the problem of predicting how much fund a Kickstarter project can receive will be the problem of predicting which range of money the project can receive.

Comparing to this work, our work is totally different, instead of predicting the range of final pledged fund, we build regression model to predict exactly the amount of money as well as the number of backers given a certain Kickstarter project.

Our contributions (or research questions) in this proposal are:

- Understand the influence of multiple factors toward the number of backers and the amount of final pledged money that a certain project can receive. We will show statistic values to illustrate for such influence.
- Given a project, we propose a model to predict how much pledged fund the creator can receive
- Finally, we Building a model to predict how many backers will fund for the project.

2. RELATED WORKS

There exists many research works on crowdfunding problem. We explain previous works based on trend of crowdfunding research.

Researchers have predicted whether the project can be successfully funded or fail. [6] collected 13,000 projects on Kickstarter and extracted 13 features from each one to develop a classifier to predict project success with 68% accuracy. [5] extends the work and show how the temporal amount of money can help improve the accuracy. [8] focused on text features of project pages and show how using phrases features to predict project success.

Another research trend tries to correlate social media activities during running fund raising campaign to project success and proposed solutions for investor recommendation problem. [7] studied how the amount of money can be affected by promotional activities on social media like Twitter. [1] used promoter network on Twitter to show the correlation between the connectivity of project promoters and project success. They also developed backer recommendation in which potential investors are suggested. [3] proposed different ways of recommending investors by using hypothesis-driven analysis of pledging behavior. [2] presented various factor influenced investor retention which allows to identify different groups of investors.

Comparing with the previous research work, we collected largest dataset consisting of more than 150k projects. Our problem is totally different comparing to existed works. That

DOI: 10.1145/1235

is, we construct statistical models that examine multiple predictive factors toward building two models: (i) one predicts the number of backers will back for the project and (ii) the another predicts the amount of pledged money that the project can receive.

3. DATASET

Our examined data set contains 151,608 Kickstarter project pages that are collected between 2009 and September 2014. Our another work [4] also used this data. In 151,608 Kickstarter project pages that we collected, there are 142,890 distinct creators.

In the following sections, we will use all 151,608 Kickstart projects pages to do analysis and to build two models to predict the number of backers who will back for a Kickstarter project as well as the amount of funding that the project can receive.

4. ANALYSIS

In statistics, linear regression is an approach for modeling the relationship between a dependent variable y and one or more explanatory variables denoted X. In this paper, there are more than one explanatory variable, the process is called multiple linear regression.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

As Figure 1 shown, only 10% of total project obtains more than 200 backers and $20,000 funding. In other words, majority of project could not achieve such a large number of backers and funding. It confirms that it is essential to success to know how many backers will fund and how much money will be collected.

Before discovering a model, we would like to see relation between major features and the number of backers, finally pledged money. In Figure 2, it gives that there is not specific linear relation between independent variables such as project duration, description length, the number of rewards, and the number of frequently asked question.

It is likely to have a shape of normal distribution in that there are more backers and participation in fund with middle of independent variables. In case of duration, there are many marks near point of "30". It has similar pattern in relation on the number of reward. This phenomenon makes hard to make model because the different dependent variables from the same independent variable. It requires to find some transformation method to fit a model.

There are similar patterns of distribution whether response variable is final pledged money or the number of backers. To understand this phenomenon, it is need to investigate relationship between them. It is obvious that there should be positive relation between the number of backers and pledged money. When more users are involved in a specific project, it is natural that the project have more fund in the last.

Figure 2 shows relation between the number of backers and final pledged money. Linear regression line is represented with the below equation.

$$FinalMoneyPledged = 71.8998 * NumBackers + 845.879 \quad (1)$$

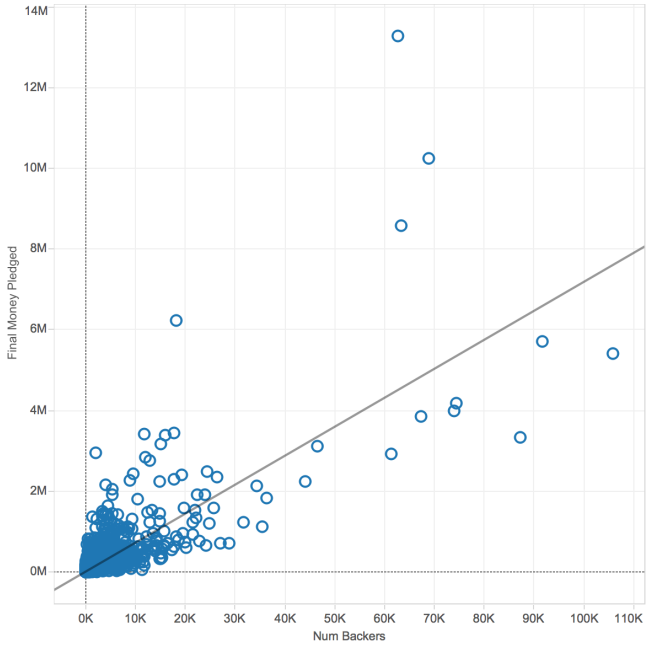


Figure 3: Scatter plot between Final pledged money and backer

With significance level 0.05, the model is acceptable because its p-value is less than 0.0001. Besides, R^2 which provides a measure of how well future outcomes are likely to be predicted by the regression model and how well the regression line ïňĄts real data.is 0.6321. It represents that this model with the number of backers explains 63.21% variance of final pledged money.

Figure 4 shows relation between final pledged money and the number of backer with previous project success rate. It represents that the more success rate in previous project, the more chance to attract backers which leads to get more financial support.This could not fully guarantee successful modeling but it point out one of features which affect project success.

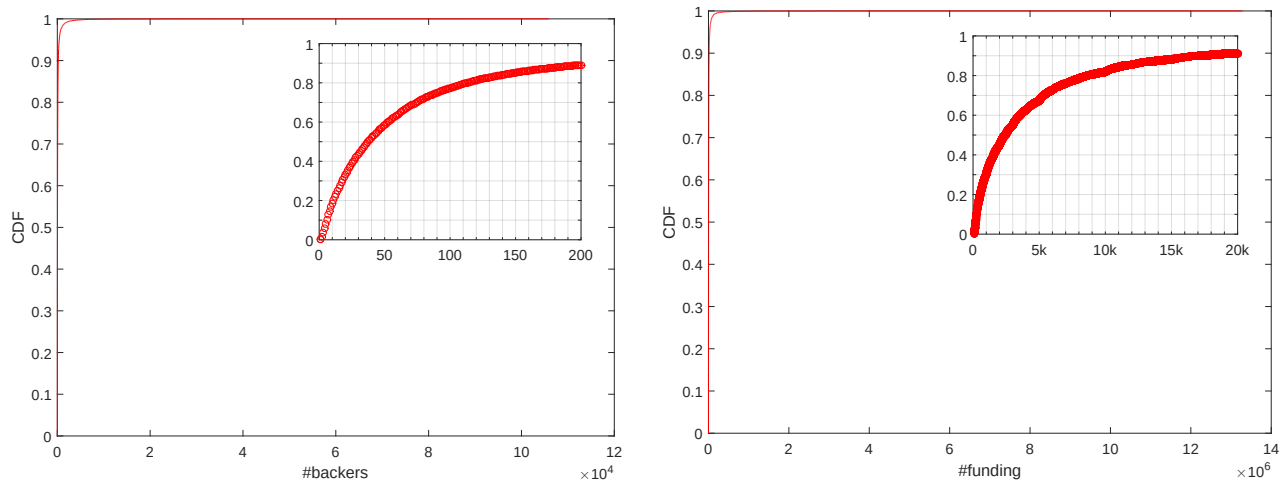
5. FEATURES

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:

5.1 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



(a) CDF of Backers Distribution

(b) CDF of Funding Distribution

Figure 1: Cumulative Distribution Function of Backers and Funding

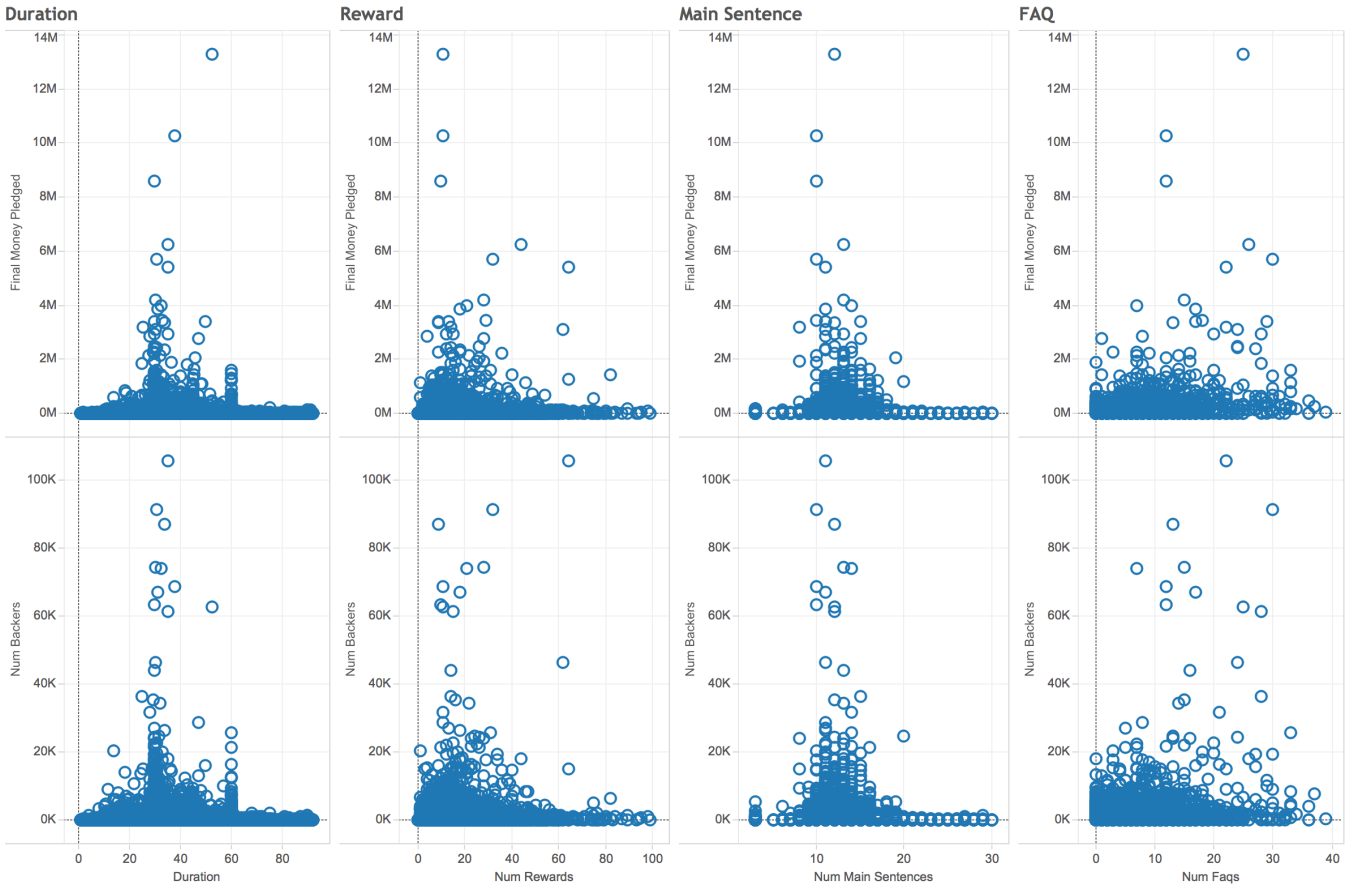


Figure 2: Scatter plot between available features and Final pledged money, backer

- 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,

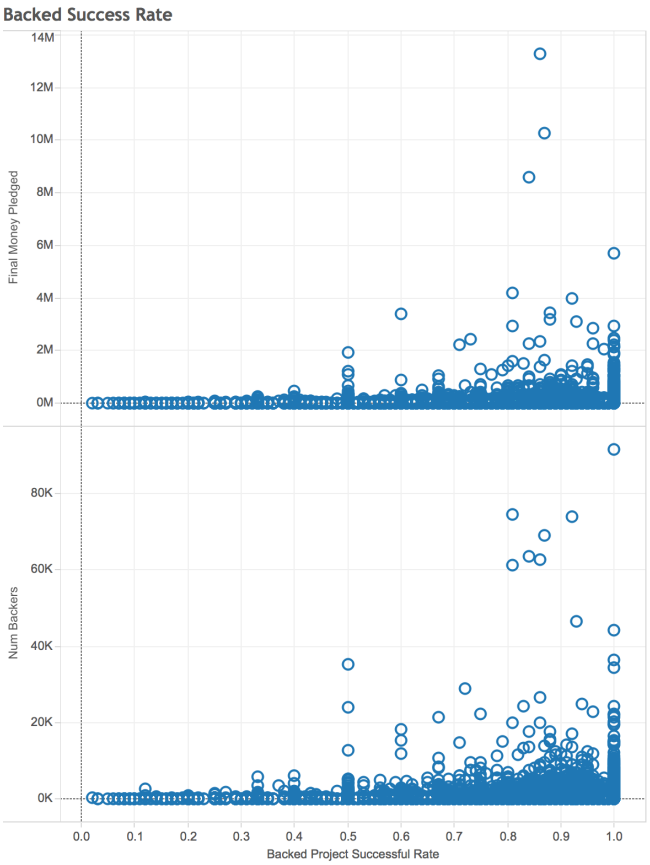


Figure 4: Scatter plot between Final pledged money and backer with previous success project rate

- 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:

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

5.2 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 [4], the more projects that creator launched, the more successful he is in raising fund.

5.3 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?

5.4 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 5.

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.

6. 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:

$$Y^{\lambda} = \beta_0 + \beta^T X + \epsilon \quad (2)$$

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:

$$\hat{Y}^{1/2} = \beta_0 + \beta^T X = X\beta \quad (3)$$

The Equation (3) can be rewritten as following:

$$\hat{Y} = (\beta_0 + \beta^T X)^2 = (X\beta)^2 \quad (4)$$

Now, the Equation (4) is a quadratic function, not a linear model. Similarly, when $\lambda = 1/3$, we have a cubic function that get through all points.

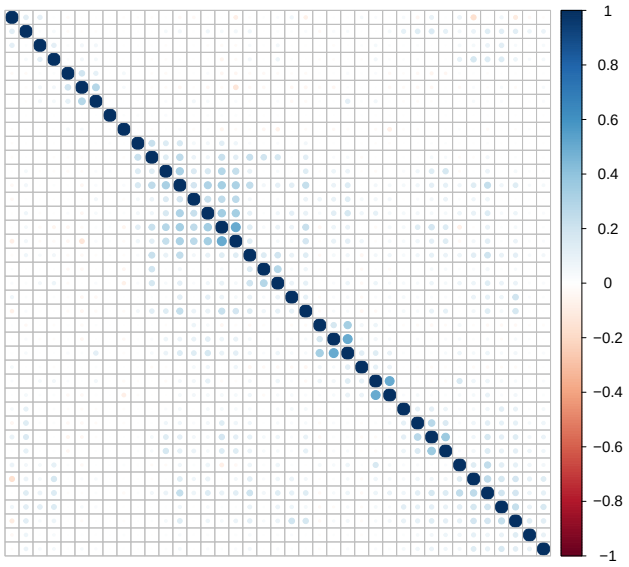


Figure 5: Correlation matrix of all features after removing highly correlated features

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:

$$\begin{aligned} f(error|\mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{error_i^2}{\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i^{\lambda} - \widehat{Y_i})^2}{\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(Y_i^{\lambda} - X\beta)^2}{\sigma^2}} \end{aligned} \tag{5}$$

Given Equation (5), the loglikelihood function of error is:

$$\begin{aligned} log(L) = -\frac{n}{2} kig(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i^{\lambda} - X\beta]^2 \\ +(\lambda - 1) \sum_{i=1}^{n} log(Y_i) \end{aligned} \tag{6}$$

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 (6).

7. 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.

7.1 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.

7.2 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 6 shows the change of loglikelihood when varying the λ value. For building the model to predict number of backers, we observed in Figure 6(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 6(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.

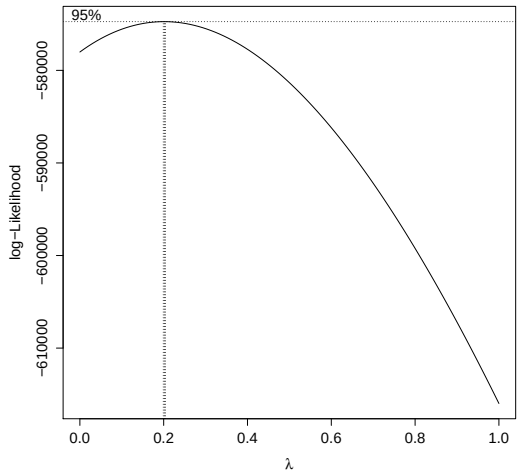
As a result, we fit the following model for predicting number of backers:

$$Y_{backers}^{0.2} = \beta_0 + \beta^T X + \epsilon \tag{7}$$

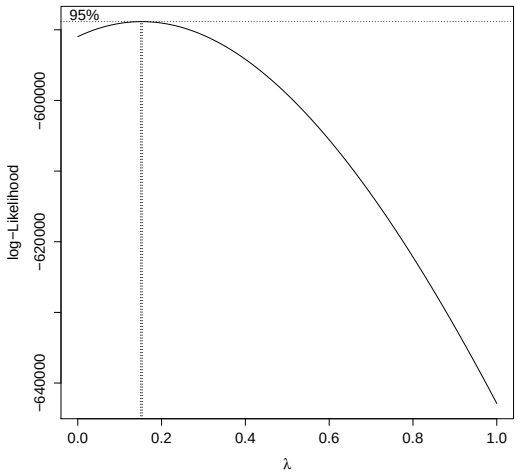
And the model for predicting the amount of funding need to be fitted is as following:

$$Y_{funding}^{0.18} = \beta_0 + \beta^T X + \epsilon \tag{8}$$

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

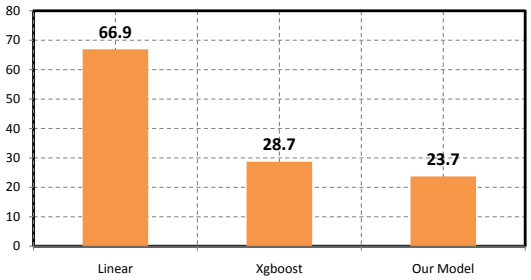


(a) Value of λ in regression model for predicting number of backers

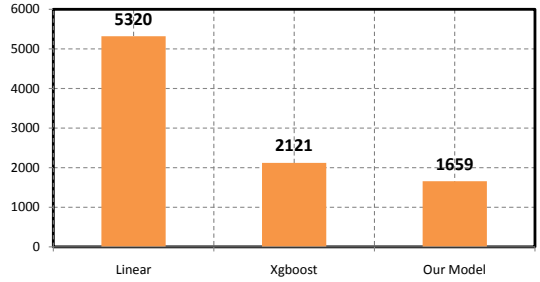


(b) Value of λ in regression model for predicting amount of funding

Figure 6: 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 7: 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

Predicting how many backers will back for a Kickstarter project: we fit our proposed model given in Equation (7). Then we compare the result with linear regression model and extreme gradient boosting for linear model. Figure 7(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 (8) 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 7(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.

8. UNDERSTANDING THE INFLUENCE OF FACTORS TOWARD THE AMOUNT OF FUNDING

Table 1 explains which features have more influence on determining the amount of funding. * represent the significance of α(*: 0.05, **: 0.01, ***: 0.001). Larger value in coefficient is more impact on prediction of the amount of funding. As table shows, the number of project comments is out-numbering other features as value of 23,540 which is a lot greater than the second ranking attribute of the number updates having 2,817. It is important to response on backers by project owners explaining communication between them is essential element of success.

Besides, some of features have negative relation with the amount of funding. The attribute which has the worst relationship with response is how many times project the creator has published in the past. This explains that backers consider the creator who made a lot in the past tends not to be successful to be funded overtime. The next following features the number of categories the creator has been made in the past. The creator did not focus on one specific area but extend its territories to several realm. Backers might sus-

Table 1: Coefficient of Features

Feature Name	Coefficient	Feature Name	Coefficient
numProjectComments ***	23539.09	cbackedCategories ***	230.262
numUpdates ***	2816.502	cbackedCategories ***	202.355
cnumFbFriends ***	1002.85	cbackedCategories ***	158.055
numRewards ***	951.701	cnumBioSentence ***	141.075
Goal ***	800.64	numFaqs ***	134.553
previousProjectSuccessRate ***	779.975	cbackedCategories *	89.731
numVideos ***	594.218	numImages ***	27.47
smogBio ***	444.779	cbackedCategories ***	13.272
cbackedCategories ***	429.861	cwebsiteCount ***	7.455
backedProjectSuccessfulRate ***	428.13	ctwitterConnected ***	-0.5
smogMain ***	341.549	cbackedCategories **	-61.033
numMainSentences ***	335.187	cyoutubeConnected ***	-183.622
Category ***	293.621	cbackedCategories **	-249.037
cbackedCategories ***	291.544	cfacebookConnected ***	-643.742
cbackedCategories ***	264.193	cbackedCategories ***	-697.825
cnumCreatorComment *	246.34	cbackedCategories ***	-847.89
cbackedCategories ***	240.012	cnumCreated ***	-1915.6

pect that the creator does not possess specialty on specific object. Backers are likely to avoid the project which project owner has no specialty.

9. 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.

10. REFERENCES

[1] *What Motivates People to Invest in Crowdfunding Projects? Recommendation using Heterogeneous Traits in Kickstarter*, 2015. Rakesh, Vineeth and Choo, Jaegul and Reddy, Chandan K.

[2] T. Althoff and J. Leskovec. Donor retention in online crowdfunding communities: A case study of donorschoose. org. In *Proceedings of the 24th International Conference on World Wide Web*, pages 34–44. International World Wide Web Conferences Steering Committee, 2015.

[3] J. An, D. Quercia, and J. Crowcroft. Recommending investors for crowdfunding projects. In *Proceedings of the 23rd international conference on World wide web*, pages 261–270. ACM, 2014.

[4] 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.

[5] V. Etter, M. Grossglauser, and P. Thiran. Launch hard or go home!: Predicting the success of kickstarter campaigns. In *COSN*, 2013.

[6] M. D. Greenberg. Public online failure with crowdfunding. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pages 333–334. ACM, 2015.

[7] C.-T. Lu, S. Xie, X. Kong, and P. S. Yu. Inferring the impacts of social media on crowdfunding. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 573–582. ACM, 2014.

[8] T. Mitra and E. Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. In *CSCW*, 2014.