SUPERVISED LEARNING MODEL FOR KICKSTARTER CAMPAIGNS WITH R MINING

R. S. Kamath[1] and R. K. Kamat[2]

[1]Department of Computer Studies, Chatrapati Shahu Institute of Business Education and Research, Kolhapur, India

[2]Department of Electronics, Shivaji University, Kolhapur

ABSTRACT

Web mediated crowd funding is a talented paradigm used by project launcher to solicit funds from backers to realize projects. Kickstarter is one such largest funding platform for creative projects. However, not all the campaigns in Kickstarter attain their funding goal and are successful. It is therefore important to know about campaigns' chances of success. As a broad goal, authors intended in extraction of the hidden knowledge from the Kickstarter campaign database and classification of these projects based on their dependency parameters. For this authors have designed a classification model for the analysis of Kickstarter campaigns by using direct information retrieved from Kickstarter URLs. This aids to identify the possibility of success of a campaign.

KEYWORDS

crowd funding; prediction; classifiers; machine learning; R systems; kickstarter

1. INTRODUCTION

Data Mining promotes distinct tools and algorithms for analyze the data patterns. Authors have explored efficiency of using machine learning algorithms for building classifiers to determine success rate of project launch. This paper explains a data mining process for investigating the relationship between the result of project launch (success, partial, failed) and a set of verticals describing the project, using the R environment and selected R packages for data analysis [10]. Reported research focused on design of classification model for Kickstarter project launch by assessing different classifiers experimentally.

The purpose of this project was to develop a system with machine learning techniques applied to Kickstarter campaigns dataset to classify projects [6, 8, 14]. To do this, authors have trained different classifiers on projects data. This approach required training data constructed by refering Kickstarter projects. This data included characteristic features of Kickstarter campaign retrieved from project URLs (https://www.kickstarter.com). This study reveals that the project properties play a vital role in predicting success.

The rest of the manuscript is organized as follows: a short theoretical background is presented in Section 2, the structural design of research framework is outlined in Section 3, Section 4 elaborates methodology carried out in data mining process, Section 5 presents classification model evaluation with experimental results, and conclusions and future research are outlined in Section 6.

DOI : 10.5121/ijitmc.2016.4102

2. PRIOR ART: EMERGING RESEARCH DIRECTIONS

In this literature review, references of the relevant work have taken and explained the same with respect to this research. This section surveys the most relevant studies carried out in this field to date. This review is supplemented by referring about 25 research papers. Some selected references for broad overview are taken here.

Chen et al. have developed a system to predict the success or failure of Kickstarter project before its completion. For this purpose they have trained support vector machine (SVM) on campaigns' data [4]. The dataset includes data retrieved from Kickstsrter projects as well as social media sources. Final classifier of this model is able to predict campaign's final outcome with 90% accuracy. The finding of this research explores that project properties are important features in determining success of a project. Etter et. al. aimed at developing a method for predicting success of Kickstarter projects by using direct information and social media [7]. They have classified the campaigns as probable success or failure based on time series of money and information retrieved from tweets and Kickstarter's projects graph. Authors have shown the importance of social feature in predicting success of projects.

Researchers from Georgia Institute of Technology Atlanta have explored the features which lead to successfully funding Kickstarter projects [13]. This study revealed that the language used in the campaign has surprising predictive power of 58.56%. Authors have explained the use of predictive phrases along with the control variables for backers and project creators to the best use of time and money. Aleyasen has developed a model to predict success of a project based on project information. To build the classifier researcher has used Kickstarter campaigns' dataset [1]. This model is able to predict success or failure of a project by 73% based on the description text. This development facilitated with a web interface which allow creator to enter campaign details and it provides feedback based on classification result.

Researchers have designed a tool for project creators to get feedback about their campaigns [12]. To accomplish this researchers have applied various machine learning classification algorithms on crowd funding projects at the time of launch. The accuracy of this tool is 68%, whether a campaign will be successful or not. The outcome of this tool is a prediction engine can be used to guide project creators. Authors have studied and analyzed the factors affecting campaign results [19]. This literature targets the project page content and usage patterns of project updates.

Semantic analysis is applied and they found discrepancies between intent of project updates and uses in practice. This analysis reveals that impact of updates rather than project details had stronger associations with campaign success. Yet another paper by Rakesh et al. explains the features determining projects' success [17]. They have expanded project features in to temporal behaviour, personal behaviour, geo-location behaviour, and social network behaviour. Using comprehensive dataset researchers have provided insights of these features and their effects on the success of campaigns. Authors have studied dynamics of Kickstarter and impact of social networks to this [11].

Literature review reveals that, development of classification model for Kickstarter campaign has been an emerging area of research in the current decade [2]. In regard to this, authors aimed at retrieving of the hidden knowledge from the Kickstarter campaigns and classification of these projects based on their features. To accomplish this authors have designed a classifiers for the analysis of Kickstarter campaigns by using direct information available online.

3. STRUCTURAL DESIGN OF RESEARCH FRAMEWORK

The structure of the paper follows the framework of a data mining process as shown in Figure 1. Reported research applied a variety of machine learning classifiers to learn the concept of online crowd funding project. A five-step procedure is followed for the design of research framework comprises:

1. Problem Definition: Analysis of relative importance of the campaign details for its success in reaching funding goal
2. Kickstarter project pages are directly parsed to get many of the project properties, and required preprocessing has done for mining purpose [18]
3. Data set preprocessed as per the requirement of machine learning algorithms
4. In supervised learning phase, classifier models are developed which associates the class variable and the explanatory variables using a training set randomly selected from the data set
5. Performance of the each classifier evaluated and selects the "best" one. It allows checking the performance of the trained classifiers against a testing set, evaluating their predictive accuracy with data not used in the training step

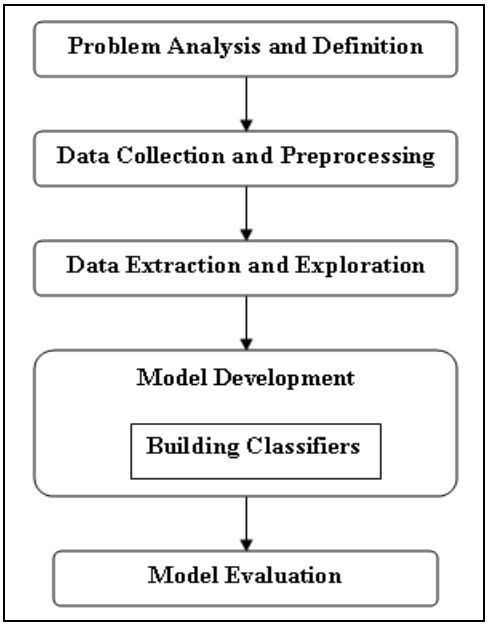

Figure 1. Data Mining Process Framework

Present research attempted to design data mining model in R, a free software environment [3]. R is a simple, but very powerful data mining and statistical data processing tool for research [16]. The mining code presented here was developed and tested using R version 3.1.2, and the corresponding scripts given here. The steps carried out in this data mining process and the R packages used are summarized in Table 1.

Table 1. Data Mining Process in R

Step	Aim	Method	R Package
1	Data Collection	Referring Kickstarter project URLs directly	----
2	Data Preprocessing	Categorical values converted in to numerical	----
3	Data Extraction and Data Exploration	Descriptive Statistics	Hmisc, ggplot
4	Building Classifiers	Decision Tree Random Forest Neural Network K-Nearest Neighbor Naïve Bayes	party, rpart randomForest Nnet Class, e1071
5	Model Selection	Confusion Matrix	Nnet

4. DATA MINING PROCESS

This section elaborates data mining process carried out in this research. Implementation of different classification models, their evaluation with experimental results are explained here.

4.1. DATA COLLECTION AND DATASET DESCRIPTION

Authors have designed dataset consists of project details retrieved directly from referring Kiststarter project URLs (www.kickstarter.com; www.quandl.com/data). Dataset provides information on over 120 project pages. The structure of Kickstarter pages includes a video, a goal, a project description, reward structure, and links to social media platforms etc. Project campaigns' main characteristics values are collected and stored in database [15]. Authors have explored a number of features from Kickstarter projects and their related data in order to performed supervised learning. Specially, we looked at the attributes given in Table 2.

Table 2. Selected attributes from Kickstarter campaigns

Attribute	Description
Project Name	Name, title of the campaign
Project URL	Kickstarter campaign URL from where details are retrieved
About Project	It explains campaign; it also takes up the majority of the space on the page.
Category	Kickstarter provides a list of categories (eg. Music, or Dance, or Video Game) and sub-categories for a new campaign. A project can be assigned either, but not both.
Sub Category	
Backers	The number of projects backed by the creator
Pledged	The amount pledged over time
Funding Goal	Amount of money (in USD) the project creator needs to raise for a project
Launch date	Date of launch of project
Funding Duration	The time limit set by the project creator, up to which the project can accept funds.
No. of	No. of updates affirming the progress of the project to existing

updates	backers and also to encourage new backers to contribute
Reward Level	Number of rewards available
video	Presence of a video explaining about project
Location	Place where the project launched
Creator	Project creator

4.2 DATA PREPROCESSING

Data Pre-processing applied for identifying the missing values, noisy data and irrelevant and redundant information from dataset. As classification algorithms works on numeric values only, categorical variables are converted in to numerical form. Kickstarter projects categories and corresponding numerical value assigned given in Table 3. Presence of video denoted as 1 where as absence as 0.

Table 3. Categorical Variable Conversion

Category	**Assignment**
Art	1
Comics	2
Craft	3
Dance	4
Design	5
Fashion	6
Film & Video	7
Food	8
Games	9
Journalism	10
Music	11
Photography	12
Publishing	13
Technology	14
Theater	15

Percentage of funding is calculated by using attributes “funding goal” and “amount pledged” [9]. It is classified in to five classes as explained in Table 4. For the purpose of supervised learning the class variable “result” is treated as dependent variable. Project density of different classes is shown in Figure 2.

Table 4. Dependent Variable “result” Description

Description (result)	**Range**	**Preprocessed value**	**No. of Observations**
0 funded	[0 – 9]	0	14
Less funded	[10 – 50]	1	28
Partial	[51 – 99]	2	38
Successful	[100 – 200]	3	25
Above goal	Above 200	4	8

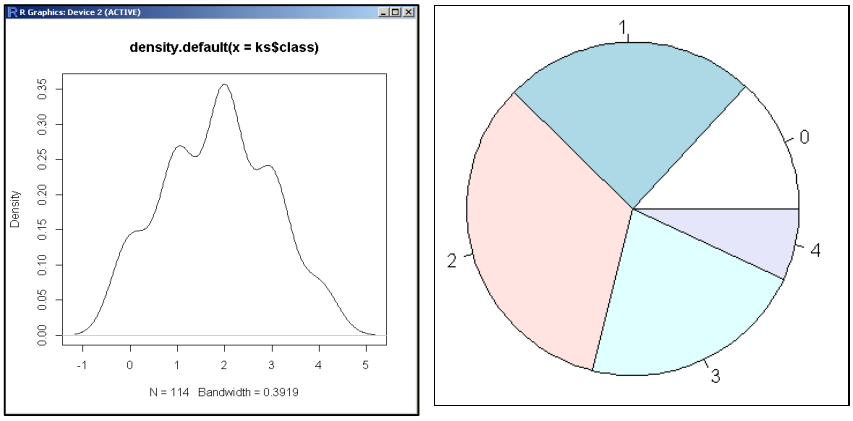

Figure 2. Class Variable Density

The following regressions were utilized to test the effects of the project presentation variables on our success measures. Authors have performed the study and analysis of various traits of Kickstarter campaigns for the selection of verticals on which project's success depends given as:

Class Variable "result" depends on: Category, Funded, Backers, Pledged, Funding goal, Duration, No. of updates, Presence of video, Reward Level

4.3. DATA EXTRACTION AND EXPLORATION – DATASET STATISTICS

Data exploration with R starts with inspecting the dimensionality, structure and data of an R object. Data set contains five classes of projects based on percentage of funding. Data frame is created by reading dataset. Results of basic statistical computation and number of observations for each type are reported by summary function as given Figure 3.

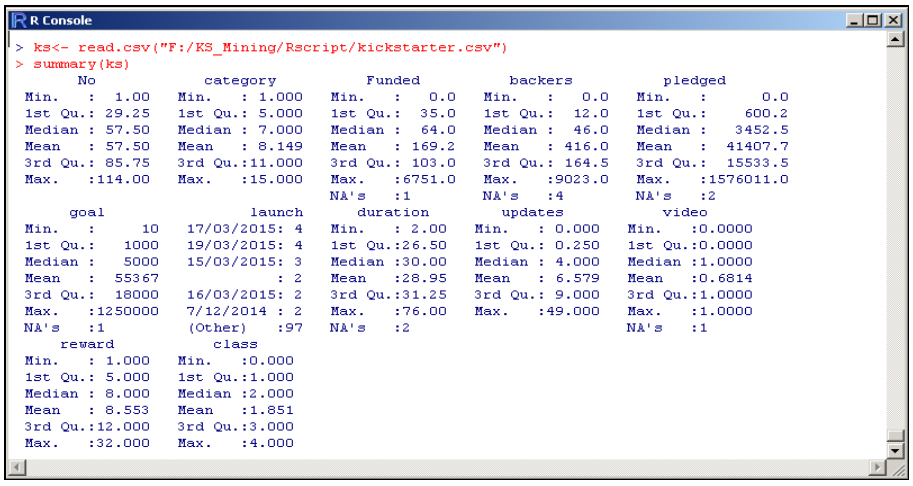

Figure 3 Basic Statistical Computations

4.4. BUILDING CLASSIFIES

Data mining model is developed by building classification rules for the target variable “result”. Authors have executed the dataset through a variety of different classification algorithms. Four classification methods during the learning step were chosen to represent a wide range of approaches in statistics and to analyze data are explained in this section.

In order to train classifiers, select their parameters and evaluate their performance, the dataset is randomly separated into 2 parts: 70% of the campaigns are selected as the training set, 30% as the validation set to evaluate the "out-of-sample" performance of the classifiers. Data set separation done in R.

4.4.1. NAÏVE BAYES CLASSIFIER

The Naïve Bayes classification is based on a probabilistic model that integrates strong independence assumptions. It can handle an arbitrary number of independent variables whether continuous or categorical. E1071, R package is installed to execute Naïve Bayes Classifier. Figure 4 explains R implementation of Naïve Bayes classifier.

```
library(class)
library(e1071)
ks<- read.csv("F:/KS_Mining/Rscript/kickstarter.csv")
ksTrain = sample(1:114,75)
ksVal = setdiff(1:114,ksTrain)
ks[ksTrain,]
ks[ksVal,]
ks$result <- factor(ks$result)
table(ks$result)
classifier<-naiveBayes(ks[ksTrain,][,1:10], ks[ksTrain,][,11])
table(predict(classifier, ks[ksTrain,][,-11]), ks[ksTrain,][,11])
classifier<-naiveBayes(ks[ksVal,][,1:10], ks[ksVal,][,11])
table(predict(classifier, ks[ksVal,][,-11]), ks[ksVal,][,11])
```

Figure 4. R implementation of Naïve Bayes classifier

4.4.2. NEURAL NETWORK

The R package NNET provides methods for using feed-forward neural networks with a single hidden layer. For the purpose of machine learning dataset separated training and validation sets. This allows validating the ANN on data that it was never trained with. The neural network requires that the records be normalized using one-of-n normalization. Input values are normalized between 0 and 1 as per the requirement of ANN. Training data is trained by “nnet” function. Generated model is tested on test data with “predict” function. R implementation of neural network is given in Figure 5.

```
ks<- read.csv("F:/KS_Mining/Rscript/ann_kickstarter.csv")
table(ks$result_cat)
ksTrain = sample(1:113,75)
ksVal = setdiff(1:113,ksTrain)
table(ks[ksTrain,]$result_cat)
table(ks[ksVal,]$result_cat)
library(nnet)
formula <- result_cat ~
            category+funded+backers+pledged+goal+duration+updates+video+reward
ksANN=nnet(formula, data=ks, subset=ksTrain, size=10, rang=0.2, decay=5e-4, maxit=200)
table(ks$result[ksTrain], predict(ksANN, ks[ksTrain,], type = "class"))
table(ks$result_cat[ksVal], predict(ksANN, ks[ksVal,], type = "class"))
predict(ksANN, ks[ksTrain,], type="class")
predict(ksANN, ks[ksVal,], type="class")
```

Figure 5. R Implementation of Neural Network classifier

4.4.3. RANDOM FOREST

The RANDOMFOREST package is used for classification by random forest classifiers. For classification the corresponding method implements Breiman's random-forest algorithm. It can also be employed for assessing proximities among data points in unsupervised mode. Figure 6 explains R implementation of Random Forest classifier.

```
ks<- read.csv("F:/KS_Mining/Rscript/kickstarter.csv")
library(randomForest)
ksTrain = sample(1:113,75)
ksVal = setdiff(1:113,ksTrain)
ks[ksTrain,]
ks[ksVal,]
table(ks[ksTrain,]$result)
table(ks[ksVal,]$result)
formula <- result ~
            category+funded+backers+pledged+goal+duration+updates+video+reward
ks_rf <- randomForest(formula, ks[ksTrain,],  ntree=1, proximity=TRUE)
table(predict(ks_rf, ks[ksTrain,]), ks[ksTrain,]$result)
table(predict(ks_rf, ks[ksVal,]), ks[ksVal,]$result)
```

Figure 6. R Implementation of Random Forest Classifier

The RPART package is used for classification by decision trees. Recursive partitioning is a fundamental tool in data mining. It explores the structure of a set of data, while developing easy to visualize decision rules for predicting a categorical value. The resulting models can be represented as binary trees. R implementation is shown in Figure 7 and corresponding tree is shown in Figure 8.

```
ks<- read.csv("F:/KS_Mining/Rscript/kickstarter.csv")
ksTrain = sample(1:113,75)
ksVal = setdiff(1:113,ksTrain)
table(ks[ksTrain,]$result)
table(ks[ksVal,]$result)
library(party)
formula <- result ~ category+funded+backers+pledged+goal+duration+updates+video+reward
ks_ctree <- ctree(formula, ks[ksTrain,])
print(ks_ctree)
plot(ks_ctree)
text(ks_ctree, use.n=T)
table(predict(ks_ctree, newdata = ks[ksTrain,]), ks[ksTrain,]$result)
table(predict(ks_ctree, newdata = ks[ksVal,]), ks[ksVal,]$result)
```

Figure 7. R Implementation of Decision Tree

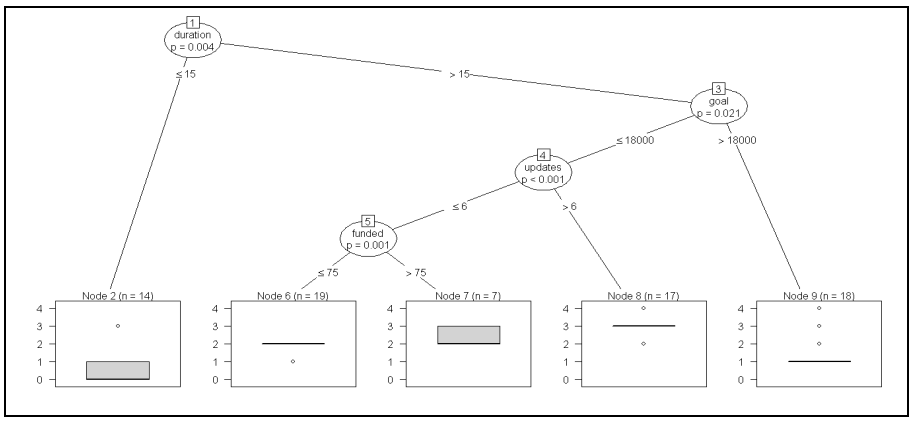

Figure 8. Decision Tree

5. MODEL EVALUATION

After developing classifiers, the criteria defined for evaluating classifiers. The performance of the four classifiers for target variable "result" can be described by the "confusion matrix", a squared contingency table. Corresponding classification model evaluation is explained in Table 5. Classification models are evaluated on the basis of their accuracy, i.e. the percentage of observations correctly classified. It is calculated as the ratio between the sum of the diagonal elements of the confusion matrix and the sample size. Method of evaluation revels that neural network is the suitable model in this case is considered as classifier for Kickstarter campaigns.

Figure 9 shows, using scatter plot, the distribution of the performances for each of the four classifiers. This graph confirms the conclusion drawn from Table 5.

Table 5. Classifier Evaluation Metrics

Classifier	Expected Result	Confusion Matrix	Accuracy
Naïve Bayes	Training Set: 0 1 2 3 4 9 21 24 15 6 Validation Set 0 1 2 3 4 5 7 14 10 2	Training Set 0 1 2 3 4 0 9 0 0 0 0 1 0 12 0 0 0 2 0 9 23 6 0 3 0 0 1 8 0 4 0 0 0 1 6 Validation Set 0 1 2 3 4 0 5 0 0 0 0 1 0 4 0 0 0 2 0 3 11 0 0 3 0 0 3 10 0 4 0 0 0 0 2	84%
Neural Network	Training Set: a b c d e 5 16 24 20 10 Validation Set a b c d e 3 9 14 8 4	Training Set a b c d e a 5 0 0 0 0 b 0 14 2 0 0 c 0 0 24 0 0 d 0 0 3 17 0 e 0 0 0 0 10 Validation Set a b c d e a 3 0 0 0 0 b 0 8 1 0 0 c 0 0 13 1 0 d 0 0 1 7 0 e 0 0 0 0 4	94%
Random Forest	Training Set: 0 1 2 3 4 10 16 29 16 4 Validation Set 0 1 2 3 4 4 11 9 9 4	Training Set 0 1 2 3 4 0 9 1 0 0 0 1 0 13 3 0 0 2 0 4 25 0 0 3 0 0 0 16 0 4 0 0 0 0 4 Validation set 0 1 2 3 4 0 4 0 0 0 0 1 0 9 2 0 0 2 0 1 8 0 0 3 0 0 0 9 0 4 0 0 0 0 4	78%
Decision Tree	Training Set 0 1 2 3 4 10 20 24 16 5 Validation set 0 1 2 3 4 4 8 14 9 3	Training Set 0 1 2 3 4 0 10 3 0 1 0 1 0 14 2 1 1 2 0 3 16 0 0 3 0 0 4 3 0 4 0 0 2 11 4 Validation set 0 1 2 3 4 0 4 2 0 0 0 1 0 3 3 1 1 2 0 3 8 0 0 3 0 0 2 4 1 4 0 0 1 4 1	52%

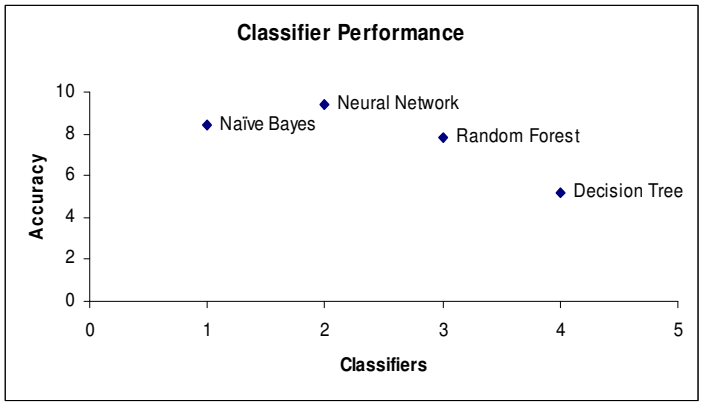

Figure 9. Scatter plot of the performance indices for classifiers

6. SIGNIFICANCE AND CONCLUSION

The present research carried out a data mining techniques implementation using different R packages. The machine learning algorithms described here are powered by scraped dataset. We were interested in design of classification model for Kickstarter campaigns. We have executed different classifiers on project dataset. Neural network is the clear winner in this contest. The results we achieved through the basic set of variables described in Table 1 are encouraging, we are able to of execute supervised learning for classification projects accurately.

The main results can be briefly summarized as follows:

1. From our analysis, we determined that the project properties play a vital role in predicting success
2. Model evaluation revels that neural network is the suitable classifier for Kickstarter campaigns.
3. Kickstarter claims that projects with a video report higher success rates than those without. Even our research proves this claim.

As research goals moving forward, the authors will be designing prediction model which will predict probability of campaign success and present reason for the probable success/failure is given. In the future, we will run more analysis on the text content of the project page.

REFERENCES

[1] Aleyasen, A. (2014) 'KickUpper: A Tool For Making Better Crowdfunding Projects'. Spring 2014.

[2] An, J., Quercia, D. and Crowcroft, J. (2014) 'Recommending Investors for Crowdfunding Projects', WWW'14, April 7–11, Seoul, Korea

[3] Carpita M., Simonetto A., Sandri M. And Zuccolotto P. (2014) 'Football Mining with R'. Data Mining Applications with R, Chapter 14, Academic Press, Elsevier,

[4] Chen, K., Jones, B., Kim, I. and Schlamp, B. n.d. KickPredict: Predicting Kickstarter Success.

[5] courses.cms.caltech.edu/cs145/2013/blue.pdf (Accessed 04 April 2015).

[6] Elizabeth, M., Gerber, Julie, S. and Kuo, P. n.d. Crowdfunding: Why People Are Motivated to Post and Fund Projects on Crowdfunding Platform. http://www.juliehui.org/wp-content/uploads/2013/04/CSCW_Crowdfunding_Final.pdf (Accessed 04 April 2015).

[7] Etter, V., Grossglauser, M., and Thiran, P. (2013) 'Predicting the Success of Kickstarter Campaigns', COSN'13, October 7–8, Boston, Massachusetts, USA.

[8] Grivenics, D., Sprovieri, D., Zwimpfer, C., Dornberger, R. and Hil, D. (2014) Crowdfunding IT Research and Business Ideas: Predicting the chances of success of a crowdfunding project. http://www.sprovieri.eu/crowdfunding_paper.pdf (Accessed 02 April 2015).

[9] Koenig, M. n.d. Make Your Kickstarter Crowdfunding Campaign Insanely Successful. http://bit.ly/FiverrKickstarterSeminar (Accessed 08 April 2015]

[10] Kuppuswamy, V. and Bayus, B. (2014) 'Crowdfunding Creative Ideas: The Dynamics of Project Backers in Kickstarter', UNC Kenan-Flagler Research Paper.

[11] Lu, C., Xie, S., Kong, X. and Yu, P. (2014) 'Inferring the Impacts of Social Media on Crowdfunding', WSDM'14, February 24–28, New York, New York, USA.

[12] Michael, D., Greenberg, Hariharan, K., Gerber, E., and Pardo, B. (2013) 'Crowdfunding Support Tools: Predicting Success & Failure' CHI'13, April 27 – May 2, Paris, France ACM.

[13] Mitra, T. (2014) 'The Language that Gets People to Give: Phrases that Predict Success on Kickstarter', Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 49–61. ACM.

[14] Mollick, E. (2014) 'The dynamics of crowdfunding: An exploratory study', Journal of Business Venturing Vol. 29 Iss. 1 pp:1–16

[15] Owano, N., (2013) Prediction model for Kickstarter tells if projects will sail, http://phys.org/news/2013-10-kickstarter.html (Accessed 27 March 2015)

[16] Pravilovic, S. (2013) 'R language in data mining techniques and statistics', American Journal of Software Engineering and Applications, Vol. 2 Iss. 1 pp 7-12
[17] Rakesh, V., Choo, J., and Reddy, C. n.d. 'What Motivates People to Invest in Crowdfunding Projects? Recommendation using Heterogeneous Traits in Kickstarter', Association for the Advancement of Artificial Intelligence, [online] www.aaai.org (Accessed 04 April 2015)
[18] Witt, N. n.d. A Kickstarter's Guide To Kickstarter: How to successfully fund your creative project', http://kickstarterguide.com/files/2012/07/A-Kickstarters-Guide.pdf (Accessed 28 March 2015)
[19] Xu, A., Yang, X., Rao, H., Fu, W., Huang, S. and Bailey, B. (2014) 'Show Me the Money! An Analysis of Project Updates during Crowdfunding Campaigns', CHI'14, April 26 – May 1, Toronto, ON, Canada. ACM 978-1-4503-2473-1/14/04