It's All Funds & Games

Predicting Kickstarter Success

Mark Giannini
University of Notre Dame
mgianni1@nd.edu

Patrick Tinsley
University of Notre Dame
ptinsley@nd.edu

Brian Tunnell
University of Notre Dame
btunnell@nd.edu

1. Introduction

1.1 Project Plan

For our semester project, we have decided to test our ability to predict whether or not a given Kickstarter campaign will be successful. In order to be deemed a success, the proposed campaign needs to meet or exceed the funding goal proposed by the initial author by a predefined deadline; anyone can contribute as long as the campaign is still active. In the context of our project, each instance in our data set has a unique project ID and fourteen features; these include project name, project description, keywords, financial goal in US dollars, the project deadline and the number of backers contributing to and supporting the project. Using sentiment analysis and other logistic regression techniques we have learned in previous classes, we plan to predict the binary final_status field, which indicates a successful project (1) or a failed attempt (0).

1.2 Data Sources

Initially, we planned to crawl the data from the Kickstarter website ourselves. However, upon browsing a plethora of Kaggle competitions, we found a pre-built data set that contains all our fields of interest. The supplied data has 108,129 rows, each corresponding to a project proposal submitted between May 2009 and May 2015. Each instance has the following features: Project ID, Name, Description, Funding Goal, Project Keywords, Disable Communication, Country, Currency, Deadline Date, Date Created, Date Launched, State Changed At, Launched At, Number of Backers and finally, the targeted response variable, Final Funding Status.

1.3 Proposed Evaluation

To evaluate our models predictive power, we plan on splitting our data into two sets. The first partition will be the training set, and it will be used to build and train our model. The second partition will be the testing set, and it will be used to validate the model. If we split the data 70%-30% respectively, the training set will have 75,690 rows, and the testing set will have 32,439 rows. By withholding a subset of the full data set, we have the power to test our final model on unseen data, which can be used to evaluate estimator performance; this technique also helps to avoid overfitting, producing a more generalized model capable of predicting the success of future Kickstarter projects.

CONF 'yy, Month d–d, 20yy, City, ST, Country.

http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn

2. Related Work

3. Problem Definition

4. Proposed Methodology

5. Data & Experiments

5.1 Data Set

5.2 Experimental Settings

5.3 Evaluation Results

6. Conclusions

A. Appendix Title

Appendix, if needed.

Acknowledgments

Acknowledgments, if needed.