CS57300 Data Mining Project Proposal

KEVIN BAIZE and YOUNGHUN LEE

.

ABSTRACT

Even though crowdfunding is one of the most popular methodologies to raise funding for a project, the reality is that not every project is able to completely reach their goal. In fact, on KickStarter, only about 188 thousand projects roughly 37% of the total projects have raised successful fundings as of 2020[1]. This fact raises an important question: which projects are able to successfully achieve their goal?. Simply put it, can project owners somehow know what features/characteristics that increases the chances of success. This is the fundamental question our team will find a solution for.

1 INTRODUCTION

Crowdfunding is the practice of funding a project or a venture by raising monetary contributions from many people across the globe[2]. There are many organization that do this but the most popular are as follows: Kickstarter, Indiegogo, DonorsChoose.org, and Patreon which hosts the crowdfunding projects on their platforms. This project will focus on data collected from Kickstarter. To give context, in order for their project to be funded they must reach their goal, and if they do then it will be distinguished as success. Kickstarter's website has raised around 5 Billion and hosted more than 500, 000 projects. What makes this project interesting is the fact that as a team we will be creating a general framework that will allow users who have a desire to put their projects on Kickstarter to have a higher probability of those projects to be funded. Taking the imagination of the project, marketing it in a way to engage the most out of other users towards their project, will ultimately give them the best chance to translate that vision towards reality.

2 PROJECT PLAN

2.1 Data

Throughout the project, we will use the ***Kickstarter Projects dataset*** posted on Kaggle [3]. The dataset includes a csv file of 15 attributes and approximately 370K instances. We are planning to consider only a subset of these attributes, such as *[name of project], [project category], [number of supporters], [fundraising goal].* and *[pledged amount].* This is mainly because other attributes are instance specific values (e.g. *[project id]*) or overlap with each other.

We have also considered using *IndieGoGo Project dataset* [4], since the dataset shares similar objective to our project. Unlike the *Kickstarter Projects dataset*, *IndieGoGo Project dataset* includes a shortened description of each project along with its title. Although the description of each project provides much more linguistic features, the *IndieGoGo Project dataset* doesn't have the information about the actual amount of fundraising goals—the dataset only indicates the pledged percentage. Since we value the actual amount of fundraising goal in determining whether the fundraising succeeds or not, our major focus would be the *Kickstarter Projects dataset.*

2.2 Algorithm Design

We aim to perform feature engineering as the first step of model implementation. Unlike other numerical attributes (fundraising goal, number of supporters, etc.), the project title includes various linguistic features we could use. For instance, we can vectorize the titles by measuring the number of words, proper nouns, capital letters, exclamation points, and so on. We could further use some contextualized pre-trained embeddings for language data such as Word2Vec [Mikolov et al. 2017] and ELMO [Peters et al. 2018], in order to capture the contextual information of each product title. Although using these approaches would be beneficial in capturing semantic of the titles, we will focus on more interpretable linguistic features to measure the correlation between features and labels. After we preprocess the data and capture the features, we will further investigate the usefulness of each feature by performing Principal Component Analysis.

To design a model, we will try out various predictive models such as Naive Bayes, Logistic Regression, SVM, as well as some ensemble models including Random Forests and XG Boost. In the case of predictive modeling, the label of each instance will be whether the pledged amount of money exceeds the goal or not. We consider a more sophisticated model by solving regression task, where the model is trained to capture the expected pledged percentage given the features.

In order to evaluate the model, we will randomly split the dataset into train, dev, and test set. For the predictive models, the model outputs whether the test instances will be able to achieve the fundraising goal in a binary manner. In this case, we measure the model performance by accuracy. For regression task, we evaluate the models by calculating the mean squared error between the predicted value and the gold value.

2.3 Timeline

- `Sep 28 - Oct 04` Literature Review
- `Oct 05 - Oct 15` Preprocessing, Feature Engineering
- `Oct 16 - Oct 18` Midterm Report
- `Oct 19 - Nov 01` Predictive Model Implementation
- `Nov 02 - Nov 15` Regression Model Implementation
- `Nov 16 - Nov 22` Visualization, Preparing Presentation
- `Nov 23 - Nov 30` Final Report
- `Dec 01 - Dec 06` Project Presentation, Buffer Area

[1]https://www.kickstarter.com/help/stats
[2]Goran Calic, "Crowdfunding", The SAGE Encyclopedia of the Internet, 2018
[3]Project URL: https://www.kaggle.com/kemical/kickstarter-projects
[4]Project URL: https://www.kaggle.com/hammsidh/indiegogo-project

REFERENCES

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405* (2017).

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).