Udacity Nanodegree Capstone Project Proposal

Neha Kumari

September 13, 2017

1 Domain background

Kickstarter is a community driven funding platform commonly known as crowdfunding. It's a pool of creative, innovative projects seeking for its way to make it life. These has become main sources of initial funding for some of the small businesses and startups who wants to launch their innovative baby to world. Consisting of millions of creators, they have the opportunity to directly connect with the public in the first go and get direct response.

Their mission being bring creative projects to life and they are true to that. Let's look some of the stats as listed on their site. Having more than 3Billion dollars pledged to projects. Approx 130,000 already successfully funded with 4 Million backers and 40 Million repeated backers. This comes to as a big motivator for getting fund for the projects.

However, the way Kickstarter works its on all or nothing basis. Having a realistic goal for a project becomes important. If a project doesn't meet the goal then the owner gets nothing. Example: if goal is of 1000 dollars, and backers funding till 999 dollars won't be a success. It has to be met exact and above. So it becomes important to understand what are the factors that will make the possibility of getting the funding much higher.

2 Problem Statement

Primary problem I am trying to solve is "Predicting whether or not a new project will be funded based on the prior data.". Secondary problem I am planning to look at is - clustering the kickstarter titles into buckets, finding out the most frequently occurring words.

3 datasets and inputs

I am planning to use already available kickstarter dataset on kaggle with below features: Dataset params

1. project-id: unique id of project
2. name: name of the project
3. desc: description of project
4. goal: the goal (amount) required for the project
5. keywords: keywords which describe project
6. disable communication: whether the project authors has disabled communication option with people donating to the project
7. country: country of project author
8. currency: currency in which goal (amount) is required
9. deadline: till this date the goal must be achieved (in Unix time-format)

10. state changed at: at this time the project status changed. Status could be successful, failed, suspended, canceled etc. (in Unix time-format)
11. created-at: at this time the project was posted on the website(in Unix time-format)
12. launched-at: at this time the project went live on the website(in Unix time-format)
13. backers-count: number of people who backed the project
14. final-status: whether the project got successfully funded (target variable – 1,0)

4 Solution Statement

For the 1st problem statement, I want to do following:

5 Solution Steps

In order to solve this problem below are the few main steps I will follow:

- Data Exploration
- Data Cleanup and Manipulation
- Feature Engineering
- Predictive Modeling Algorithm using a supervised learning model
- Result Analysis

For the 2nd problem statement, I want to do following:

6 Solution Steps

In order to solve this problem below are the few main steps I will follow:

- Title Cleanup
- Vectorizing the titles
- apply clustering using kMeans
- creating buckets

7 Benchmark Model

I didn't find a benchmarking for this. However, done by some folks on kaggle.

8 evaluation metrics

For the predictive model, using cross validation and accuracy to find out how well model is performing.

9 Project Design

Project design will pretty much follow the solution steps with main building blocks like as below: