Machine Learning Engineer Nanodegree ¶

Capstone Proposal

Sifan Liu
July, 2018

Proposal

Domain Background

Kickstarter is one of the leading crowd-sourcing platform focusing on early-stage funding for creative entrepreneurship. Research has shown 90% of successful projects remianed ongoing ventures . However, 1 in 2 of the launched projects would fail. What are the key factors that determine the success of a campaign? Business dynamism has been declining across U.S., as a policy analyst, I am personlly interested to see if the analysis could shed some linght on potential policy intervention.

Problem Statement

This project aims to explore what factors could best explain a successful fundraising campaign on Kickstarter. According to Kickstarter (https://www.kickstarter.com/help/stats (https://www.kickstarter.com/help/stats)), success rates vary across 20% - 60% for different categories. Within each category, there are significant variations in terms of the initial amount seeked, the way the project was presented/described, charateristics of the launcher (gender, city, etc.), or even the day the project went live.

Datasets and Inputs

Webrobots started to web-scrap kickstarter data every month since 2014 and they've made all datasets publicly available in json and csv files (https://webrobots.io/kickstarter-datasets/ (https://webrobots.io/kickstarter-datasets/)). Although Kickstarter limits the amount of historic projects you can get in a single run, the active and latest projects are always included. The datasets include the following fields:

1. Photo: the feature of the project
2. Blurb: the short description of the project
3. Goal: amount of USD asked
4. Pledged: amount of USD raised
5. Status: failed/canceled/successful/suspended/live
6. Slurg: Headline of the project
7. Currency
8. Deadline: the project ended
9. Created time: the project was created
10. Launched time: the project went live
11. Number of backers
12. USD pledged: converted pledged amount
13. Creater profile: a link to html page including Name, Bio, Profile picture
14. Creater location: town, state, country

15. Category and sub-category

The raw dataset includes various types of inputs including natural language (description and headline), images (feature photo, profile picture), and both discrete and continuous data.

Solution Statement

This aims to build a supervised classification model,in which the goal is to classify a Kickstarter project into failue or success, using inputs from the the individual campaign, as well as geographic-spefic fixed effects including median household income, unemployment rate, etc.

Benchmark Model

The benchmark model used data on all projects that finished between: 6/18/2012 and 11/9/2012 (Greenberg et al. 2013). The attributes are summarized in the table below:

1. Goal in dollars of the project
2. Project category (eg. Music, or Dance, or Video Game)
3. Number of rewards available
4. Length of project in Days
5. Connected to twitter Boolean
6. Video present Boolean
7. Connected to Facebook Boolean
8. Number of facebook friends
9. Number of twitter followers
10. Sentiment (pos, neg, or neutral)
11. Number of sentences in project description
12. Outcome variable Boolean

The authors evaluated the performance of radial basis, polynomial and sigmoid kernel functions with varying costs for support vector machines. The also tested decision tree models with AdaBoosting. The best performing model were able to predict the success of a crowdfunding project with 68% accuracy.

Evaluation Metrics

The data will be split into training and testing set. I will evaluate the models by comparing their performance on the testing set. The key evaluation metrics is accuracy score, measures out of all projects, how many could the model correctly classify as success.

Project Design

Data cleaning and wrangling

A significant part of the analysis will be data cleaning. Key steps would include:

- check for duplicated projects
- parsing dates
- crosswalk location to Metropolitan Statistical Areas and merge with census data.
- train an algorithm to determin the gender of campaign owner based on the names (and potentially, profile pictures)

- using natural language processing to classify campagin title and description
- check for missing values
- scale and normalizae data

Explotary Data Analysis

In this step, I will run some basic summary statistics to understand the data better

- distribution of each variables
- correlatio matrix among variables
- summarise baseline successful rates by key variables of interest: category/gender/location

Feature selection

I will use Lasso and PCA to select a subset of the variables before fitting the model.

Modeling:

I will use following models to fit the dataset. It's important to understand what factors are contributing the most to the final result, so the interpretbility of the final model is of interest.

- Logistic model
- Classification tree

Evaluation

The models will be evaluated using the evaluation matrix discused above

Discussion

Based on the results of the evaluation metrics, I will select the one with best predicing power, and look into and discuss how that knowledge could be used to encourage entrepreneurship