Udacity Machine Learning Engineer Nanodegree

August 18, 2018

1 Capstone Project

Sifan Liu

1.1 I. Definition

1.1.1 Project Overview

New ideas need financial resources to succeed, and many seek their initial support from crowdfunding. Research has shown 90% of successful projects on crowdfunding websites remianed ongoing ventures *(Mollick and Kuppuswamy 2014)*. 1 in 2 of the launched projects however, would fail to raise the amount they need. While the quality of the idea itself may dicate the patterns of success, funders rely on the web-presentation of the project to identify the actual quality of the idea. Therefore, many other project attributes could also potentially determine the crowdfunding success.

1.1.2 Problem Statement

Kickstarter is one of the leading crowd-sourcing platform focusing on early-stage funding for creative entrepreneurship. According to Kickstarter , 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 and described, charateristics of the launcher (gender, location, etc.), the size of the amount to raise, or even the planned length of the campaign. This project aims to develop a tool to help predict a project's likelihood of success upon its launch.

This project aims to using these attributes to predict if a campaign would fail or succeed. Creators could leverage the predictions to make changes, and potentially increase their likelihood of success.

1.1.3 Metrics

I will split the data into training and testing set, first train the models on the training set and then 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. In this specific case, the prediction doesn't have to be either high precision or high recall, so accuracy score or F1-score would suffice.

1.2 II. Analysis

1.2.1 Data Exploration

I use the universe of the Kickstarter projects since its launch in 2009 to February, 2018. Webrobots started to web-scrap kickstarter data every month since 2014 and they've made all datasets publicly available in json and csv files.

One caveat is that Kickstarter limits the amount of historic projects you can get in a single run, which means the lastest dataset might not include all previous projects. According to Webrobots, the active and latest projects are always included in the crawl, so one way to get around this barrier is to combine different datasets at mutiple dates. To do that, I selected one dataset each year from 2014, and merge the four datasets and identif the union as the final data to analyze.

The crawl result contains information from the project webpage, and not every column is relevant for this analysis. For example, information that is updated *after* the project launch, such as total amount of USD raised, total number of backers, etc. should not be included as these varibles are not observed at the project launch.

There are a couple charateristics about the dataset needs to be noted:

1. The string variables are analyzed using Natural Language Processing. Specifically, I use sentiment analysis to determine the subjectivity and polarity of Blurb and Slug. I also use a pre-trained gender classificaiton model to assign gender information (female/male/mostly female/mostly male/unknown/andy) to creators based on their first names. Noting that Kickstarter creators could be organizations, rather than individuals. Those organization names are usually classified as "unknown".

2. This project wants to predict whether the 'status' of the campaign would be "failed" or "successful" when the campaign ends, given the initial attributes of a project. Noting that "Live" projects are still in the fundraising stage and their final status is subject to change. Projects are "suspended" when the Kickstarter team uncovers evidence that it is in violation of Kickstarter's rules, and projects are "canceled" when the creators want to cease the fundraising process for any reason. The decision to suspend or cancel projects are usually independent of the project attributes, and is not what this model aims to predict. Therefore, projects with status other than "successful" or "failed" will be excluded from the dataset.

3. Time difference between "created" and "launched" is a proxy of how much time the creator spent on preparing for the campaign; time difference between "deadline" and "launched" indicates the lenghth of fundraising decided by the creator. I will calculate both measures and include them in the model.

Key variables of interest are described as follows:
[Table 1. Variables]

Variables	Type	Descriptions
Blurb	Str	A short description of the project
Goal	Num	Amount of USD asked
Status	Factor	Five status: failed/canceled/successful/suspended/live
Slurg	Str	Headline of the project
Deadline	DateTime	Time the project ended
Created	DateTime	Time the project was created

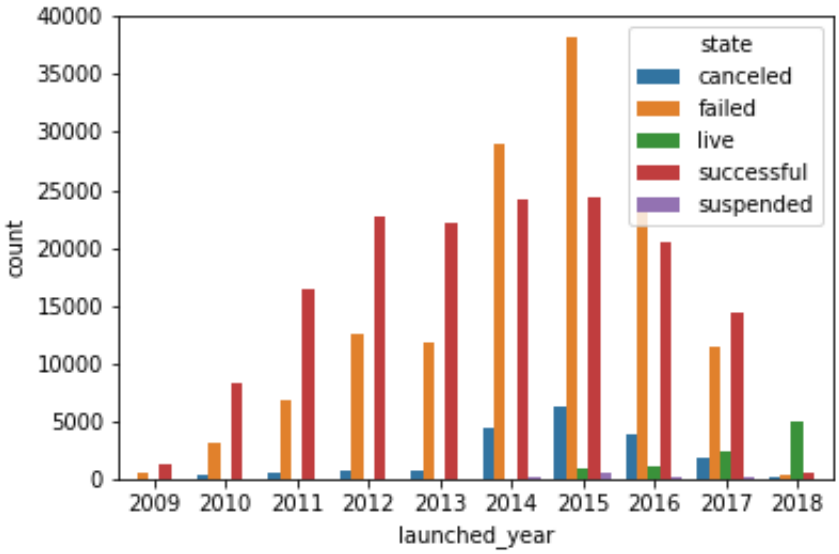

Figure 1. Total number of projects by project status

Variables	Type	Descriptions
Launched	DateTime	Time the project went live
Creator	Str	First name and last name of the creator
Location	Factor	Town, state, country of the creator
Category	Factor	15 unique categories

The descriptive statistics of numeric values are displayed in the table below:
[Table 2. Summary Statistics]

Stats	goal	life_days	prep_days	slug_plor	slug_subj	blurb_plor	blurb_subj
count	241717	241717	241717	241717	241717	241717	241717
mean	3.8e+04	33.6	43.8	0.04	0.17	0.14	0.40
std	1.0e+06	12.8	114.9	0.19	0.28	0.26	0.29
min	1.0e-02	1.0	0.0	-1.0	0.00	-1.0	0.0
25%	2.0e+03	30.0	2.0	0.00	0.00	0.00	0.16
50%	5.0e+03	30.0	10.0	0.00	0.00	0.10	0.40
75%	1.3e+04	35.0	34.0	0.00	0.30	0.29	0.59
max	1.0e+08	91.0	2313.0	1.00	1.00	1.00	1.00

1.2.2 Exploratory Visualization

First I examined the distribution of outcome to predict, project status. *Figure 1* shows the number of projects by their status in each year. It's clear from the plot that the 'canceled' and 'suspended' projects represent a very small portion of the datasets in each year, and removing them from the analysis should not cause selection bias.

While pledged amount is not included in the predictors as it is not observable at the launch

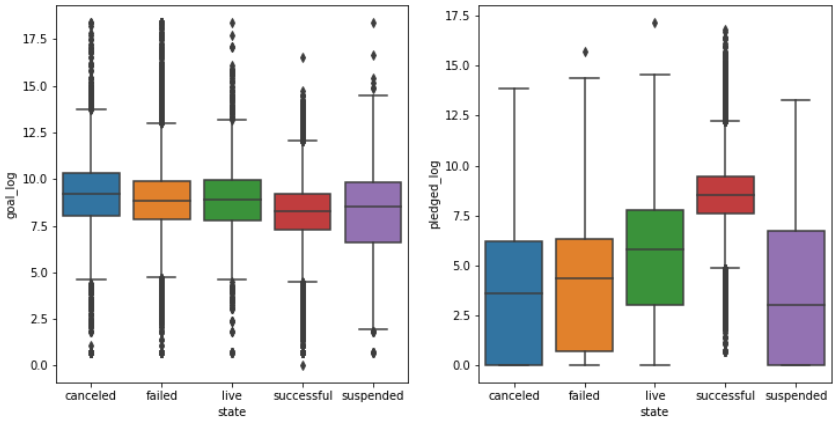

Figure 2. Goal and pledged amount by project status

state	canceled	failed	live	successful	suspended
launched_year					
2010	0.02	0.27	0	0.71	0
2011	0.03	0.28	0	0.69	0
2009	0.04	0.29	0	0.67	0
2013	0.02	0.34	0	0.64	0
2012	0.02	0.35	0	0.63	0
2017	0.06	0.38	0.08	0.48	0
2014	0.08	0.5	0	0.42	0
2016	0.08	0.48	0.02	0.42	0
2015	0.09	0.54	0.01	0.35	0.01
2018	0.02	0.05	0.83	0.09	0

Figure 3. Share of projects by project status each year

of the project,it's important to understand its relationship with the project status and project goal. *Figure 2* illustrates that successful projects raised more while asked for less, compared to failed projects.

Next I explored patterns of successful projects, whether the rate of success varies by category, gender and location. The crosstab suggests that projects with certain attributes tend to have higher rate of success.

Figure 3 shows that projects launched in earlier years are more likely to succeed on Kickstarter. This could have several explanations. First, the platform is more competitive as Kickstarter become more popular over the year. As shown in Figure one, the total number of successful projects remained stable, but failed projects has steadily increased. Second, early adoptors of the platform are usually enthusits of crowdsourcing and therefore, might be more serious about their project quality, and make more effort to promote the campaign.

Figure 4 shows that on average, names perceived as female are more likely to succeed. This is consistent with (Marom, Robb, and Sade 2016) that crowdfunding reduces the barriers of female entrepreneurs to raise capital. The authors find that while female creators made up about 35% of the project leaders, their rates of success are higher than male creators, even after controlling for category and goal amount.

The third table showed significant variance in successful rates for different category of projects - dance, theatre, comics and design on average have a successful rate over 60% while fashion, journalism and technology have a successful rate below 30%.

state	canceled	failed	live	successful	suspended
gender					
mostly_female	0.05	0.35	0.02	0.58	0
female	0.05	0.38	0.02	0.54	0
andy	0.05	0.39	0.03	0.52	0.01
unknown	0.06	0.39	0.03	0.51	0
mostly_male	0.07	0.45	0.03	0.46	0.01
male	0.06	0.48	0.03	0.43	0

Figure 4. Share of projects by project status for type of gender

state	canceled	failed	live	successful	suspended
category_broad					
dance	0.02	0.15	0.01	0.83	0
theater	0.04	0.19	0.01	0.76	0
comics	0.05	0.22	0.03	0.7	0
design	0.03	0.17	0.12	0.67	0
film & video	0.04	0.36	0.02	0.58	0
music	0.05	0.39	0.01	0.54	0
games	0.05	0.37	0.06	0.51	0.01
photography	0.07	0.43	0.01	0.48	0.01
art	0.07	0.45	0.02	0.46	0
publishing	0.03	0.54	0.02	0.4	0
crafts	0.07	0.51	0.02	0.39	0.01
food	0.08	0.56	0.02	0.34	0.01
fashion	0.1	0.57	0.04	0.29	0.01
journalism	0.08	0.65	0.01	0.26	0.01
technology	0.13	0.59	0.03	0.25	0.01

Figure 5. Share of projects by project status for each category

1.2.3 Algorithms and Techniques

This is a supervised learning problem, and the prediction is a binary outcome, with successful projects coded as "1" and failed projects coded as "0". The exploratory visualization also indicates that the underlying relationship is rougly linear, so logistic algorithms would be suitable for this question. The logit output can also be interpreted as a probability, a nice feature to have for predicting the successful rate.

Two additional emsemble models are also utilized to compare the performance:

1. Random forest: good for both linear and non-linear models.
2. Boosted Tree: good for large feature sets and non-linear models

Scaling and normalization techniques are applied to numerical variables before fitting the models. Categorical variables are converted to dummies using one-hot-encoding.

1.2.4 Benchmark

The naive predictor, in which we predict every campaign to be successful, has an accuracy score of 0.5415, and an F-1 score of 0.5962. This will serve as the main benchmark model to evaluate the model performance.

Another benchmark model is a similar project from a per-reviewed study *(Greenberg et al. 2013)*. The author used data on all Kickstarter projects that finished between: 6/18/2012 and 11/9/2012. The attributes used as predictors are slightly different, summarized 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.

1.3 III. Methodology

1.3.1 Data Preprocessing

One major component of the data processing is utilizing text analysis to convert project title and descriptions to numeric values. For this task, I use sentiment analysis from TextBlob package. Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

An example from the dataset is shown below:

Last year, I completed my first novel, The Circumstance of Marriage. Now I need your help to get an editor to get it published!

The assigned sentiment value is:

Sentiment(polarity=0.15625, subjectivity=0.19999999999999998)

After text analysis, key steps to process the data are:

1. **Duplicated projects**: the dataset is composed of data from 4 independent crawls on 2015-12-17, 2016-06-15, 2017-02-15 and 2018-02-15. Therefore, duplicated projects are checked and removed.
2. **Outliers**: A data point with a feature that is beyond an outlier step outside of the IQR (1.5 times) for that feature is considered abnormal. Observations with at least two abnormal features are removed from the analysis (1.4% of the entire observations)
3. **Data normalization**: Log transformation is applied to all numeric varaibles to achieve a normalized distribution
4. **Data scaling**: although logistic models and decision trees are not affected by feature scaling, SVM and CNN are still susceptible. min-max scaler is applied to all nummeric variables.

1.3.2 Implementation

Three models were trained on the prepocesed training data (190,609 observations). Training time, accuracy score and F-1 score are stored to compare the performance of different models.

Then I use GridSearch to search the optimal parameters to tune the performance of Random Forest model. A list of tuning parameters is shown below:

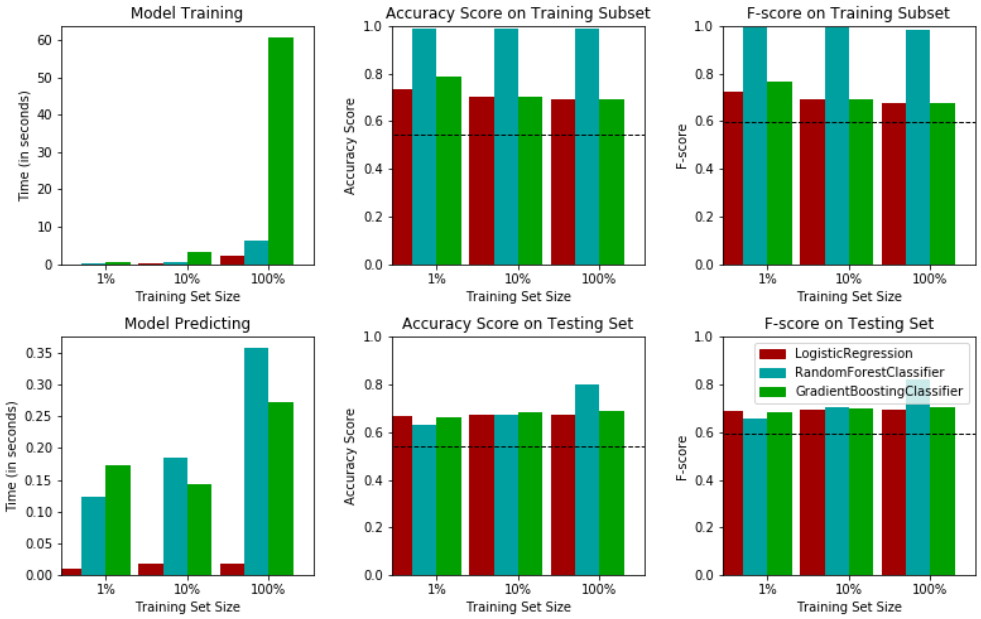

Figure 6. Predicting time and accuracy score

- **n_estimators**: [10, 50, 100, 500]
- **min_samples_split**: [2, 5, 10]
- **min_samples_leaf**: [1, 5, 10]

This improved the accuracy score from 0.8018 to 0.8201, and F-1 score from 0.8218 to 0.8247

Finally, I use feature importance to look for the features with most predictive power in the model.

1.3.3 Refinement

For the text analysis, I initially considered classifying the sentence structure of the headline and use structure type as input. This involves calculating pair-wise similarity matrix for each tokenized titles, and then using K-means to find clusters. However, this approach failed to identify meaningful types of sentence structures, and I decided to use sentimental analysis instead.

For the training models, I also considered Support Vector Machine and Conventional Neural Network. However, the training time for 190k observations turns out to be computationally expensive.

1.4 IV. Results

1.4.1 Model Evaluation and Validation

20% of the dataset is held for testing. The model is trained on the rest of the data, with 190,609 obseravations. Figure 6 shows the time and score of three models for 1%, 10% and 100% of the sample respectively.

I chose the RandomForest model as the final model because it has the highest accuracy score on testing set, within a reasonable training time.

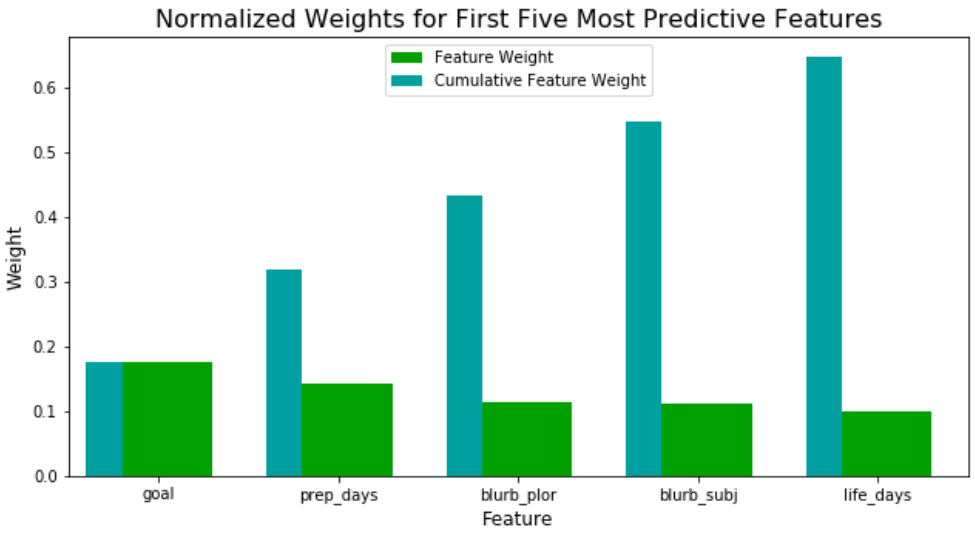

Figure 7. Feature importance of Random Forest algorithm

The normalized weights for five most predictive features in the random forests are Goal, Days spent on preparing for the campaign, Polarity of the campaign description, Subjectivity of the campgain description and lastly, Days to raise money.

1.4.2 Justification

All three models outperfom the accuracy score of the navie predictor (accuracy score of 0.5415, and an F-1 score of 0.5962). The logistic model achieved an accuracy score of 69%, which is very similar to the score published by the peer reviewed study.

Random forest outperforms other models when trained on full set by a large margin.

1.5 V. Conclusion

1.5.1 Reflection

Overall, I think a 82% accuracy rate meets the expectation of giving people guidance on whether their chance of getting funded successfully. Attributes that are most important to the change of success, namely campaign goal and project description, are easy fix that people can implement to increase their likelihood of success. How well this model translates to other crowdsourcing sites such as Indiegogo however, is in question as different sites might attract very different funders who may have different preferences over certain attributes.

One issue I realized with the Random Forest model (or other machine learning model) is that the method is designed to solve the problem of prediction, not parameter estimation. However, if the model predicts a low success rate of a campaign, it would be helpful to also indicate what attributes are holding back the performance, and how to improve the success rate. While Random Forest does produce featuer importance, it lacks estimation of standard errors on the coefficients.

As an experiment, I applied Logistic Regression from the stats package which gives parameter estimations on the same training set. The regression result is shown below, and Figure 8 shows the coefficients of each feature.

[Table 3. Regression results]

var	coef	std err	z	P> z	[0.025	0.975]
goal	-6.2287	0.069	-90.612	0.000	-6.363	-6.094
life_days	-1.3034	0.053	-24.362	0.000	-1.408	-1.199
prep_days	1.8486	0.026	71.725	0.000	1.798	1.899
slug_plor	1.6692	0.126	13.203	0.000	1.421	1.917
blurb_plor	1.4147	0.121	11.644	0.000	1.177	1.653
blurb_subj	-0.0235	0.017	-1.362	0.173	-0.057	0.010
gender_female	0.3356	0.050	6.649	0.000	0.237	0.434
gender_male	-0.1434	0.050	-2.893	0.004	-0.241	-0.046
gender_mostly_female	0.4490	0.055	8.212	0.000	0.342	0.556
gender_mostly_male	-0.0688	0.054	-1.268	0.205	-0.175	0.038
gender_unknown	0.2531	0.050	5.064	0.000	0.155	0.351
launch_month_2	0.1245	0.025	4.917	0.000	0.075	0.174
launch_month_3	0.1706	0.025	6.881	0.000	0.122	0.219
launch_month_4	0.0847	0.025	3.402	0.001	0.036	0.134
launch_month_5	0.0757	0.025	3.068	0.002	0.027	0.124
launch_month_6	0.0408	0.025	1.641	0.101	-0.008	0.090
launch_month_7	-0.1587	0.024	-6.528	0.000	-0.206	-0.111
launch_month_8	-0.1098	0.025	-4.431	0.000	-0.158	-0.061
launch_month_9	0.0656	0.025	2.615	0.009	0.016	0.115
launch_month_10	0.1270	0.025	5.134	0.000	0.078	0.175
launch_month_11	0.1004	0.025	3.986	0.000	0.051	0.150
launch_month_12	-0.0517	0.028	-1.876	0.061	-0.106	0.002
category_broad_comics	1.2398	0.035	35.788	0.000	1.172	1.308
category_broad_crafts	-0.4674	0.038	-12.316	0.000	-0.542	-0.393
category_broad_dance	1.9088	0.054	35.618	0.000	1.804	2.014
category_broad_design	1.6045	0.034	47.021	0.000	1.538	1.671
category_broad_fashion	-0.7141	0.029	-24.605	0.000	-0.771	-0.657
category_broad_film & video	0.8775	0.021	41.897	0.000	0.836	0.919
category_broad_food	-0.1135	0.026	-4.339	0.000	-0.165	-0.062
category_broad_games	0.6375	0.026	24.880	0.000	0.587	0.688
category_broad_journalism	-0.7694	0.045	-16.971	0.000	-0.858	-0.681
category_broad_music	0.4412	0.021	21.500	0.000	0.401	0.481
category_broad_photography	0.3203	0.035	9.179	0.000	0.252	0.389
category_broad_publishing	-0.2539	0.022	-11.777	0.000	-0.296	-0.212
category_broad_technology	-0.4170	0.026	-16.310	0.000	-0.467	-0.367
category_broad_theater	1.6380	0.036	45.054	0.000	1.567	1.709

This show the magnitude as well as the direction of each parameters: if you want to improve your rate of success, you should consider a smaller goal, go for a shorter fundraising timeframe, spend more time preparing for the campaign, and make sure your project description is both positive and subjective.

1.5.2 Improvement

There are a couple areas to improve this project:

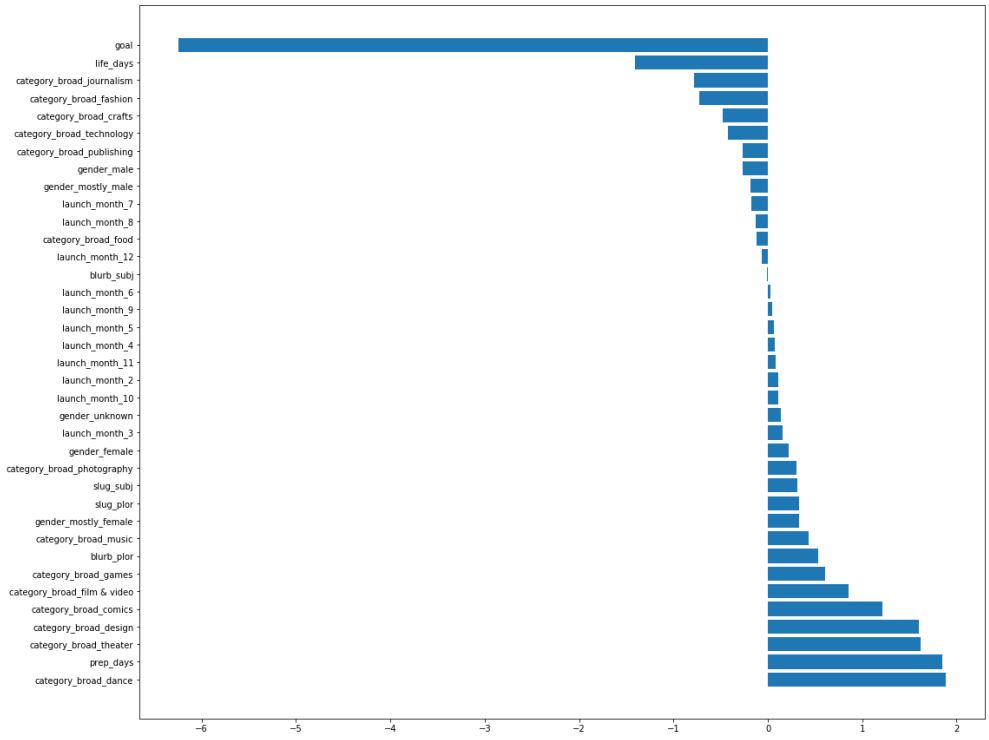

Figure 8. Logistic regression coefficients

1. Image analysis: another potential attribute to include would be the characterstics of creator's profile image: would presenting a human face increase the chance of getting funded? Does the hue of the image matter? This however, requires significant storage and computing power to implement using Conventional Neural Network.

2. Feature selection: PCA is not recommended for datasets containing a mix of continuous and categorical variables. Research has suggested using multiple correspondence analysis for mixed data types. I need more understandings of the techniques before implementing it.

3. Sentiment analysis: Ideally I want to train the sentiment analysis classifier on the Kickstarter dataset, which would be more relevant.

2 Reference

- Greenberg, Michael, Bryan Pardo, Karthic Hariharan, and Elizabeth Gerber. 2013. "Crowdfunding Support Tools: Predicting Success & Failure." In , 1815–20. doi:10.1145/2468356.2468682.
- Marom, Dan, Alicia Robb, and Orly Sade. 2016. "Gender Dynamics in Crowdfunding (Kickstarter): Evidence on Entrepreneurs, Investors, Deals and Taste-Based Discrimination."
- Mollick, Ethan R., and Venkat Kuppuswamy. 2014. "After the Campaign: Outcomes of Crowdfunding."