KICK STARTER

Course: COMM 414
Instructor: Andy Chen

Student:
61994570 Tianling Wang
60105335 Casey Zhou
77798353 Shuang Wu

1. Business Understanding:

1.1 Background

Kickstarter is a public-benefit corporation, which maintains a global crowdfunding platform focused on creativity and merchandising. On Kickstarter, you can raise crowdfunding for your startup, or back other crowdfunding projects that you consider as promising. Kickstarter was launched in 2009, with the mission to "help bring creative projects to life". In 2010, Kickstarter was named as one of the Best Invention of 2010 by Time. Since 2011, Kickstarter gradually opened projects based in United Kingdom, Canada, Australia, China, Japan, Singapore and Other different countries. Until 2018, Kickstarter has already become one of the most popular crowdfunding platforms in the world, and received more than $4 billion in pledges from 15 million backers for 154,000 creative projects.

However, despite the popularity, the success rate of all projects on Kickstarter is only 36.55%, according to the official data. After glancing over the website, we found that successful projects appear to share certain similarities. Therefore, our group decided to dig into these projects, and attempted to analyze the relationship between the final status of the project and its attributes.

1.2 Problem Identification

Firstly, we want to identify what factors on the Kickstarter website presentation will influence the success of project. Once the project is launched, the fund raiser needs to set a series of features on Kickstarter website, including 'goal', 'reward level', 'duration', etc. We propose that these features could have influence on the probability of success. Secondly, we want to build a model to predict whether a certain project will succeed or fail. Finally, we would like to discuss how fund raisers can increase their fund raised by modifying these attributes.

This business problem can be classified into two aspects——business goal and data mining goal: The business goal of our research is to provide useful information for the stakeholders mentioned below. The success criteria for business goal is that, stakeholders can benefit from our outputs: Knowing what factors on the Kickstarter website presentation will influence the status of project and how to increase the money they can raise, fund raisers will do better when design their presentation on Kickstarter. Meanwhile, with the model predicting whether a Kickstarter will succeed or not, backers and Kickstarter company can benefit respectively from rewards and commissions. On the other hand, the data mining goal is identifying the factors that are highly correlated with final status (the target variable), predicting the final status, and increasing the money raised (funded percentage) by changing independent variables. The success criteria for data mining goal is 80%+ predictive accuracy.

1.3 Motivation

1.3.1 Benefits to Stakeholders

For fund raisers and entrepreneurs, our research is meaningful, since the entrepreneurial environment is not hopeful——82 percent of the new businesses fail to survive because of cash flow problems; 27 percent of new businesses surveyed by the NSBA claimed that they weren't able to receive the funding they needed. Even on Kickstarter, the success rate is only 36.55%. Therefore, investigating the relationship between the final status and the attributes of these projects can help fund raisers to better design their project presentation on Kickstarter, increasing the probability to succeed.

For backers who invest their money into these projects, they are always seeking for promising projects. Since if the projects succeed, they will get reward corresponding to the money they invested. If the projects fail, their money will be returned. But if so, they'd better put these money into banks rather than invest them into failed projects, due to the time value of money. Therefore, the results of our research would also be valuable to them.

Another stakeholder is Kickstarter Company. Not only can our research provide useful operational information, it can also help Kickstarter increase its profits, given that Kickstarter will charge for 3%-5% commissions if a project succeeds.

1.3.2 Differentiation

In addition to the benefits brought to the stakeholders, another reason for us to conduct this research is, our project is differentiated from any other previous projects and thus worthy to be done. Some researches, like *what Sarchak posted on Github*[1], have explored the text mining of keywords and descriptions, and their relationship with the final status of each Kickstarter project. Other projects, like the one conducted by Ann Rajaram on Kaggle[2], have applied decision tree to explore whether the goal, currency, deadline, state, and country features can provide a good prediction of the project's final status. However, until now, all researches in this field failed to consider the important presentation and interaction factors, including updates, awards, comments, FAQs, and video condition of each project. That's what exactly our research attempts to probe into. Moreover, in order to provide a comprehensive and accurate result, our research will apply multiple models—logistic regression, decision tree, and linear regression to get a synthetic prediction.

2. Data Understanding & Preparation:

2.1 Data Understanding

After comparing a large number of datasets about Kickstarter, we finally choose this one from Kaggle uploaded by Peter Joseph Arienza a year ago[3], because this is the one with the most comprehensive variables, especially detailed reward setting data, which we assume will be closely correlated with success rate. This dataset contains 45814 data points and 17 variables which are project id, name, url, category, subcategory, location, status (successful, failed, live, canceled, or suspended), goal (target amount), pledged (actual pledged amount), funded percentage (pledged / goal), backers, funded date, levels (the number of reward levels), reward levels (the reward amount in each level, separated by comma), updates (the number of information or activity updates created by project initiator), comments and duration.

Based on our business problem, the prediction subject in our logistics regression and random forest models is the status, whether the project is successful or not, which is a non-numeric binary outcome. The prediction subject in our linear regression model is the funded percentage, which is a continuous outcome. All variables are included in our model, except for project id, name and url.

2.2 Data Preparation

In general, the data is clean, but still occasionally there are some missing data and messy data and besides we need to change some of the data into more analyzable format.

2.2.1 Data Cleaning Using Excel

In step one, we found some data points whose contents are unmatched with titles. Therefore, we deleted all of them. In step two, we used '*status*' attribute to filter all the data points with status of '*cancelled*', '*suspended*' and '*live*', and deleted all of them, since these data don't contribute to our research goals, which intensely related to '*successful*' or '*failed*'. In step three, we deleted all the data points whose contents cannot be identified in Python, and get the final dataset of 40553 data points.

2.2.2 Data Formatting Using Python

Originally, the dataset contains reward levels in one cell, and it's extremely hard for us to conduct data mining. Therefore, we firstly split the values using excel, and then upload them on python, to calculate their averages, standard deviation, and correlation of variation.

project	iname	url	reward levels
39409	WHILE THE	http://www.kicksta	$25,$50,$100,$250,$500,$1,000,$2,500
126581	Educatior	http://www.kicksta	$1,$5,$10,$25,$50

[1] https://github.com/sarchak/MachineLearningNotebooks/blob/master/Kickstarted%20Success%20Prediction.ipynb

[2] https://www.kaggle.com/anu2analytics/how-to-raise-money-on-kickstarter/notebook

[3] https://www.kaggle.com/parienza/kickstarter

```
data.insert(0, 'Average', np.mean(data, axis=1))
```

```
data
```

	Average	reward levels split	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
0	632.142857	25.0	50.0	100.0	250.0	500.0	1000.0
1	18.200000	1.0	5.0	10.0	25.0	50.0	NaN

```
data.insert(0, 'Std', np.std(data, axis=1))
data
```

	Std	reward levels split	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
0	826.212776	25.0	50.0	100.0	250.0	500.0	1000.0
1	17.859451	1.0	5.0	10.0	25.0	50.0	NaN

After that, we add these new attributes using Excel: average, standard deviation (std), and correlation of variation (CV) into the original dataset, use these three attributes to capture the characteristics of rewards, and delete the reward column. Then, we re-upload the dataset to python.

For the convenience of modeling, we remove all space in the column names and replace with underscore.

```
data.columns = data.columns.str.replace(' ', '_')
```

```
data['funded_date'].str.split(' ')
```

```
data = pd.merge(data, pd.DataFrame(data['funded_date'].str.split(' ',expand=True)),
                how='left', left_index=True, right_index=True)
```

```
data.rename(columns={0:'weekday', 1:'date', 2:'month', 3:'year'}, inplace = True)
```

```
data = data.drop(columns=['funded_date', 4, 5])
```

```
data['weekday'] = data.weekday.apply(lambda x: x.replace(',',''))
```

The data format of "funded_date" variable includes all information about weekday, date, month, year and time. For instance, "Fri, 19 Aug 2011 19:28:17 -0000". We hypothesize that the pattern of weekday, date, month and year may have different influences on the success rate, so we separate these information by space into 5 columns. Since the specific time in a day when the project is initiated may have little impact and is hard to measure, so we drop these two columns.

Then we drop the rows where elements are missing.

```
data.dropna(axis=0, how='any', inplace=True)
```

After cleaning all data, we export the cleaned dataset for EDA analysis. But for modeling, we still need to get some dummy variables.

```
data = pd.get_dummies(data, columns=['status'],drop_first=True)
```

```
data = pd.get_dummies(data, columns=['category'],drop_first=True)
```

```
data = pd.get_dummies(data, columns=['subcategory'],drop_first=True)
```

```
data = pd.get_dummies(data, columns=['location'],drop_first=True)
```

```
data = pd.get_dummies(data, columns=['weekday'],drop_first=True)
```

```
data = pd.get_dummies(data, columns=['month'],drop_first=True)
```

For categorical variables which includes status (successful or failed), category (14 categories), subcategory (51 subcategories), location (4849 locations), weekday (Sun - Mon) and month (Jan - Dec), we divide them to separate columns (one less than the number of unique values in that variable) taking 0 or 1 indicating which category it belongs to. This file is exported for modeling use.

2.3 Exploratory Data Analysis

Using data cleaned, we want to have a better understanding of the features and relationships of key variables.

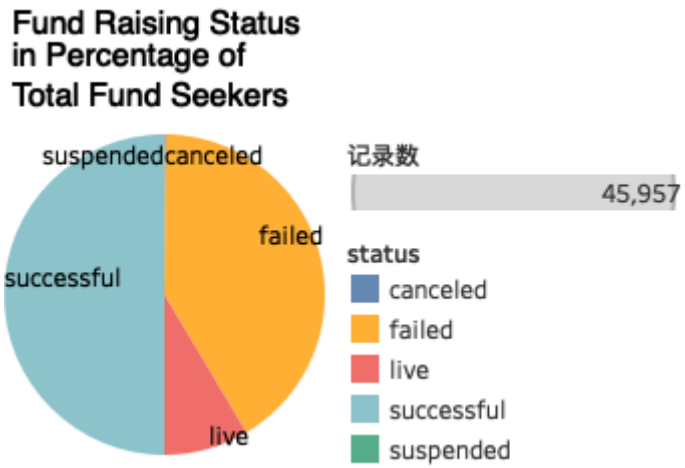


Figure 1: Pie chart of fund raising status

Our target variable is the fund-raising status. Using the original dataset, we found half of the projects are successful, less than half are failed and a small minority are live, canceled and suspended which are removed in modeling.

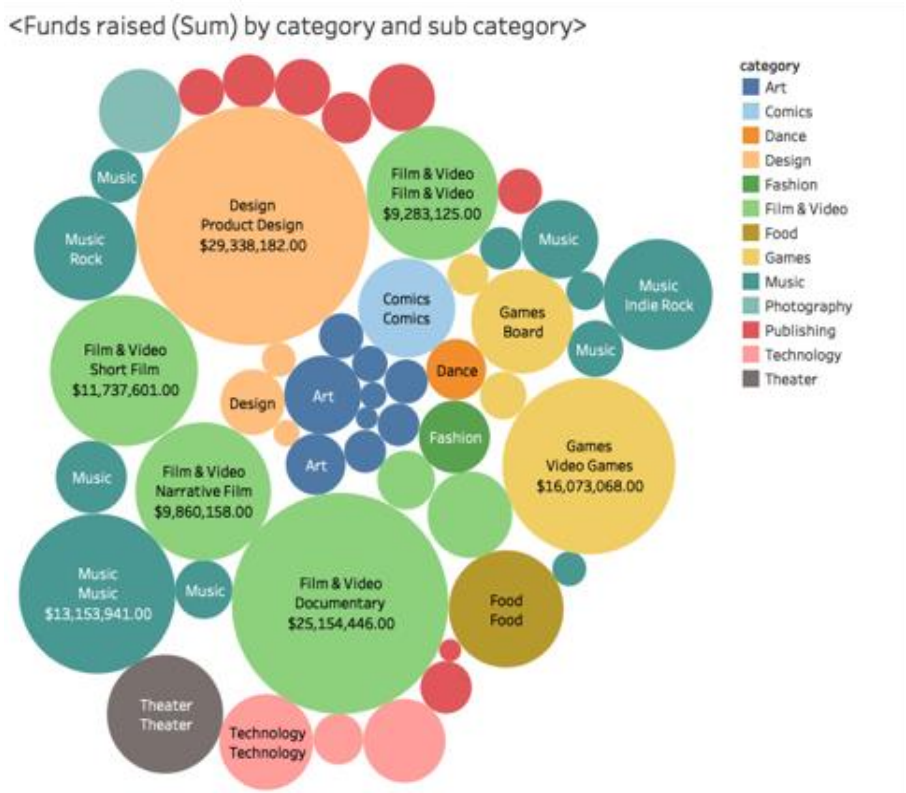


Figure 2: Bubble Plot of Category

Using bubble chart in Tableau, we calculate and visualize the total amount raised by category and subcategory, which shows Product Design, Documentary Film and Video Games gain the most funds.

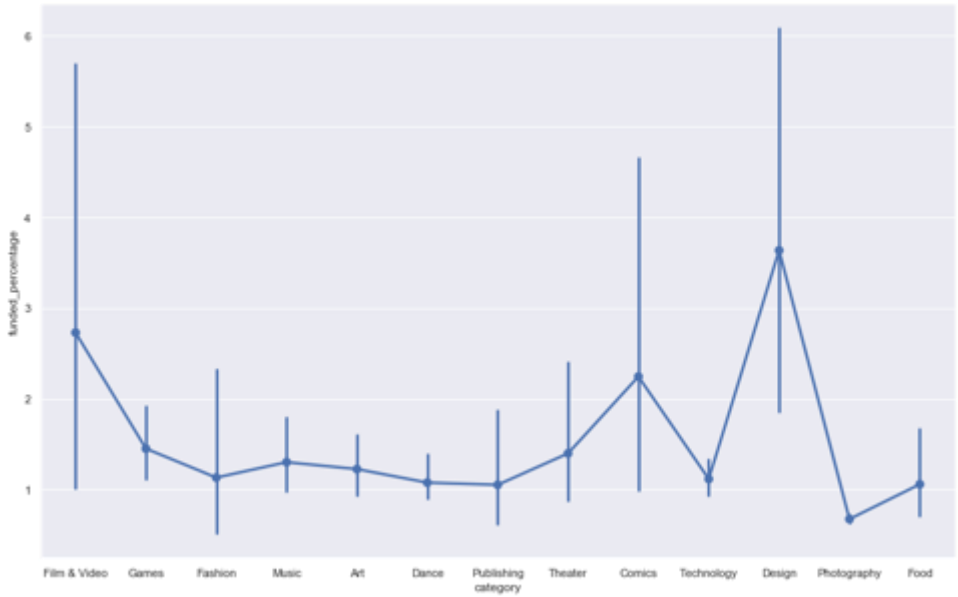


Figure 3: Point Plot of Category

For funded percentage, we use pointplot in seaborn to show the distribution. On average, Design projects have the highest pledged/goal ratio, which shows people are willing to fund design projects even if it has reached the goal, probably because getting rewards of design product are attractive.

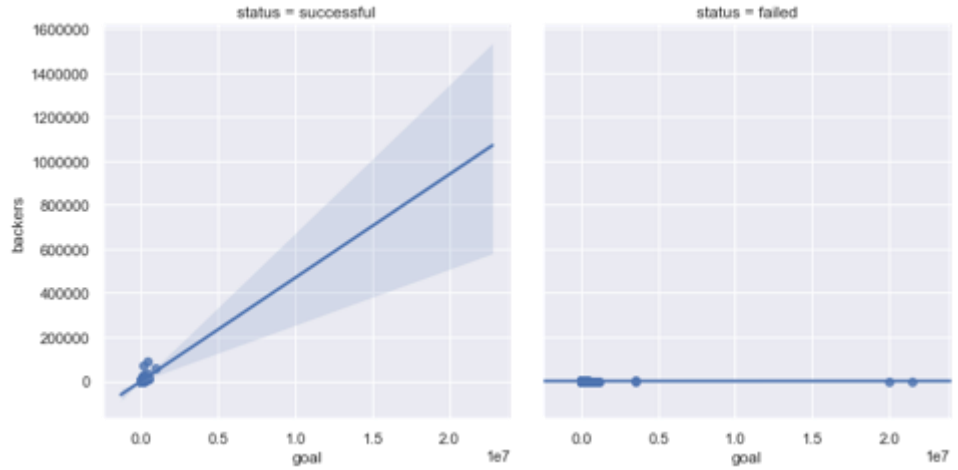


Figure 4: Implot of Fund Raising Status

For relational analysis, we plot the lmplot of goal and backers categorized by funding status. There's positive correlation if the project is successful, while there's no significant relationship if it is failed. That is because failed projects have few backers no matter how much the goal is, while successful projects have enough backers meeting the specific goal of each.

	Unnamed	project_id	goal	pledged	funded_percentage	backers	levels	updates	comments	duration	Average	Std	CV	date	year	FAQ	video	length(s)
Unnamed: 0	1.000	1.000	-0.004	-0.001	-0.003	0.005	-0.004	-0.003	-0.004	0.002	-0.008	-0.008	-0.008	-0.006	0.002	0.002	0.025	0.008
project_id	1.000	1.000	-0.004	-0.001	-0.003	0.005	-0.004	-0.003	-0.004	0.002	-0.008	-0.008	-0.008	-0.006	0.002	0.002	0.025	0.008
goal	-0.004	-0.004	1.000	0.026	-0.001	0.035	0.016	-0.001	0.033	0.034	0.069	0.071	0.026	-0.000	0.003	0.017	0.004	0.007
pledged	-0.001	-0.001	0.026	1.000	0.010	0.830	0.070	0.097	0.558	0.002	0.060	0.068	0.048	-0.000	0.024	0.104	0.013	0.019
funded_percentage	-0.003	-0.003	-0.001	0.010	1.000	0.006	0.006	0.014	0.006	0.002	0.002	0.004	0.008	0.008	-0.005	0.002	-0.003	-0.006
backers	0.005	0.005	0.035	0.830	0.006	1.000	0.075	0.110	0.706	-0.006	0.053	0.062	0.048	-0.002	0.029	0.147	0.018	0.023
levels	-0.004	-0.004	0.016	0.070	0.006	0.075	1.000	0.262	0.086	0.040	0.301	0.358	0.578	0.001	0.092	0.147	0.199	0.196
updates	-0.003	-0.003	-0.001	0.097	0.014	0.110	0.262	1.000	0.101	0.053	0.070	0.091	0.170	-0.014	-0.074	0.202	0.137	0.116
comments	-0.004	-0.004	0.033	0.558	0.006	0.706	0.086	0.101	1.000	-0.006	0.033	0.039	0.034	0.002	0.022	0.137	-0.000	0.013
duration	0.002	0.002	0.034	0.002	0.002	-0.006	0.040	0.053	-0.006	1.000	0.101	0.102	0.058	-0.017	-0.200	-0.021	-0.019	0.034
Average	-0.008	-0.008	0.069	0.060	0.002	0.053	0.301	0.070	0.033	0.101	1.000	0.950	0.461	0.007	0.042	0.061	0.109	0.156
Std	-0.008	-0.008	0.071	0.068	0.004	0.062	0.358	0.091	0.039	0.102	0.950	1.000	0.585	0.006	0.048	0.078	0.123	0.166
CV	-0.008	-0.008	0.026	0.048	0.008	0.048	0.578	0.170	0.034	0.058	0.461	0.585	1.000	0.005	0.045	0.102	0.212	0.193
date	-0.006	-0.006	-0.000	-0.000	0.008	-0.002	0.001	-0.014	0.002	-0.017	0.007	0.006	0.005	1.000	0.006	0.007	0.006	-0.004
year	0.002	0.002	0.003	0.024	-0.005	0.029	0.092	-0.074	0.022	-0.200	0.042	0.048	0.045	0.006	1.000	0.107	0.069	0.058
FAQ	0.002	0.002	0.017	0.104	0.002	0.147	0.147	0.202	0.137	-0.021	0.061	0.078	0.102	0.007	0.107	1.000	0.086	0.063
video	0.025	0.025	0.004	0.013	-0.003	0.018	0.199	0.137	-0.000	-0.019	0.109	0.123	0.212	0.006	0.069	0.086	1.000	0.610
length(s)	0.008	0.008	0.007	0.019	-0.006	0.023	0.196	0.116	0.013	0.034	0.156	0.166	0.193	-0.004	0.058	0.063	0.610	1.000

Figure 5: Correlation Table

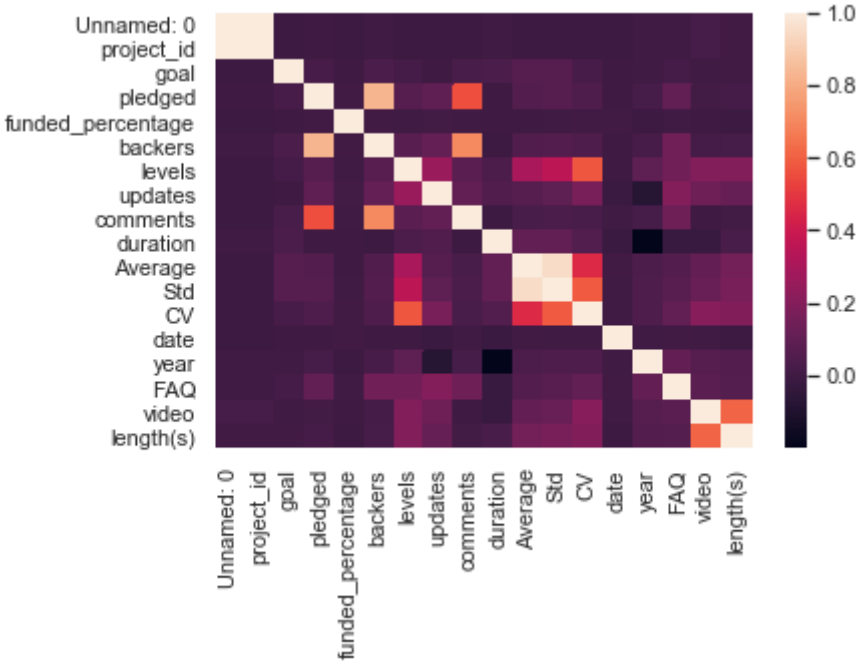


Figure 6: Heat Map

To identify the potential correlations between all numeric variables, we plot the correlation table and the heat map. While the majority of them show weak correlation, a few of them have significant Pearson' correlation coefficient.

(1) *pledged* and *backers*: r=0.83; *backers* and *comments*: r=0.706. Intuitively, the more comments it gets suggest that the more backers have interest in the project, and the more backers involved, the more money being pledged. We didn't include *pledged* and *backers* as our predictor variables, since they are the final outcomes of the project, which have similar attributes as target variables.

(2) *Average* and *Std*: r=0.95; *Std* and *CV*: r=0.59; *Average* and *CV*: r=0.46. These variables (average, standard deviation, correlation of variation) capture the distribution of the money pledged for each reward level of the project. Statistically, standard deviation is greatly dependent on the average level, so correlation of variation is introduced to reduce that effect. So we use *Average* and *CV* in our models to better represent the reward level features and at the same time avoid multicollinearity.

(3) *video* and *length(s)*: r=0.81. Since *video* is a binary variable with 0 for no video and 1 for with video and *length(s)* is a continuous variable which is also 0 if there's no video presented, they have strong correlations. We will only use length(s) as our predictor variable, since it is a more accurate record capturing the video's feature.

More exploratory data analysis graphs are presented in the Appendix (Figure 1 - Figure 13) for better analyzing and visualizing the dataset.

2.4 Additional Data Crawling:

While screening through the raw data from Kaggle, there are several significant website features not included in the dataset, such as whether there is video and the length of the video, how many FAQs and the creators' profile. Moreover, since the data is created in 2012, we want to update some afterwards data to the ongoing project in dataset and cross-validating the raw data.

Therefore, we firstly try to use beautiful-soap package in Python to crawl data from Kickstarter website. However, since there are various types of website put the video information in different divisions, it increases the difficulty of obtaining complete data. Due to time limitation, we adopt another web crawler application called bazhuayu web crawler[4]. By using this crawler, we crawl the video length, number of FAQ from over 40,000 kickstarter website.

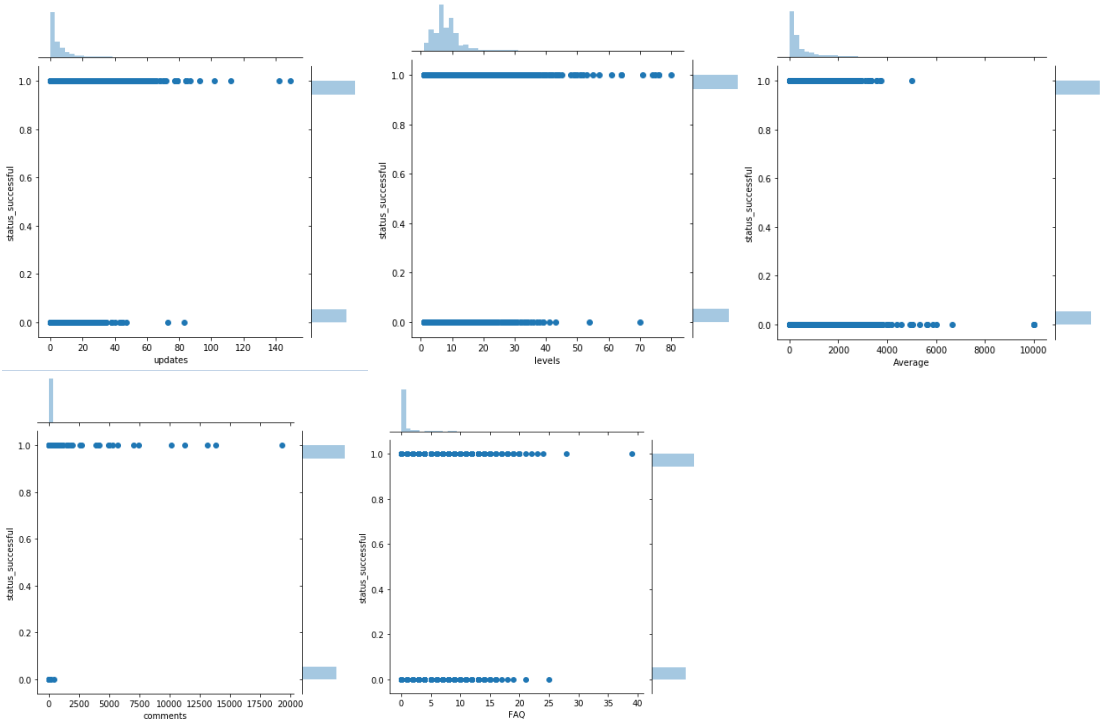
3. Modeling:

3.1 Logistic Regression

We apply logistic regression to detect what factors on the Kickstarter website presentation will influence the final status of project.

Firstly, since we use logistic regression to explain the phenomena rather than only predict the results, the multi-collinearity needs to be avoided. According to the correlation table, we found that the *Average* and *Std* (standard deviation) are highly correlated (0.95), so do *video* and *lengths* (0.81). Therefore, in the following modeling, we will avoid using them together. Moreover, we decided not to include *backers* and *pledged* in our model, since fund raisers cannot control these two variables when initiating the project. They are the final outcomes of the project highly correlated with final status.

Then, to select variables for our initial model, joint plots are drawn to see the relationship between each predictor and the target variable.



[4] http://www.bazhuayu.com

From joint plots above, we found that for successful and failed projects, these features are significantly different. We proposed that, these five variables would have influence on the final status of a project. Therefore, we include them as predictor variables in our initial logistic regression.

Trail 1: Predictor variables: 'updates', 'levels', 'Average', 'comments', 'FAQ'

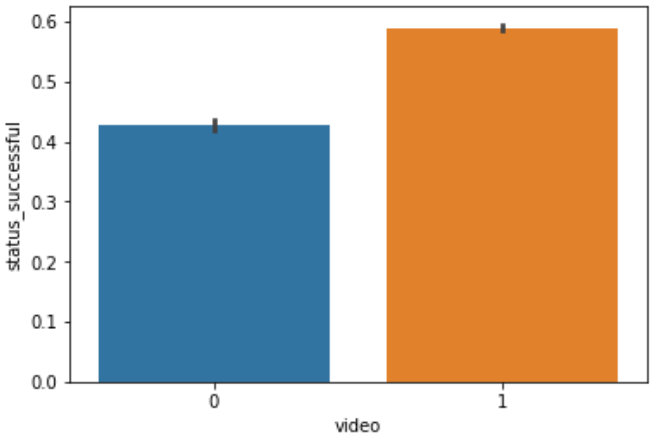
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
updates	0.2667	0.0054	49.2057	0.0000	0.2561	0.2773
levels	-0.0460	0.0027	-17.1963	0.0000	-0.0512	-0.0408
Average	-0.0011	0.0000	-28.7289	0.0000	-0.0011	-0.0010
comments	0.1393	0.0053	26.1077	0.0000	0.1288	0.1497
FAQ	-0.1353	0.0112	-12.0566	0.0000	-0.1572	-0.1133

```
confusion matrix:
 [[4504  923]
 [1853 4886]]
```

```
Model: Logistic Regression
Accuracy: 0.771823113595
Precision: 0.841108624548
Recall: 0.725033387743
F1-score: 0.778769525024
```

All p-values are less than 5% indicating that these five variables are significant.

We further tried to include other possible variables. From the bar plot of video and final status, we found that, whether the website contains promotion videos will result a different successful rate as a whole. If the website presents a video, it will have a higher probability to succeed. Therefore, we will include *'video'* in our second trail.



Trail 2: Predictor variables: 'updates', 'levels', 'Average', 'comments', 'FAQ', 'video'

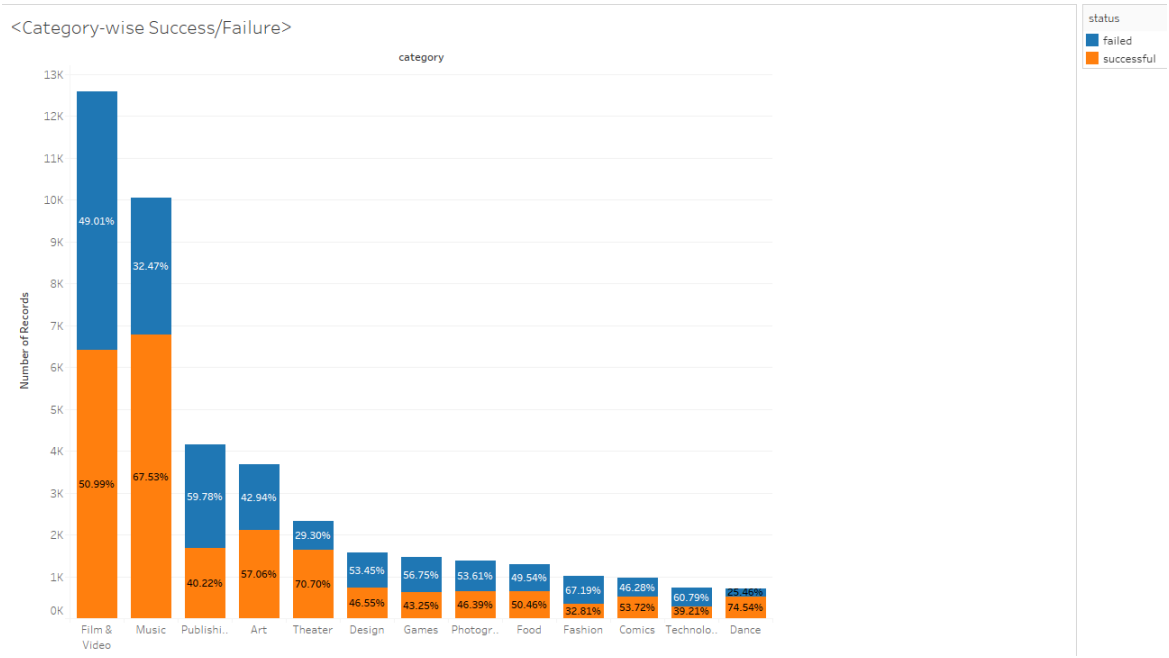
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
updates	0.2683	0.0055	49.1033	0.0000	0.2576	0.2790
levels	-0.0410	0.0033	-12.2710	0.0000	-0.0476	-0.0345
Average	-0.0011	0.0000	-28.5342	0.0000	-0.0011	-0.0010
comments	0.1398	0.0053	26.1423	0.0000	0.1293	0.1503
FAQ	-0.1344	0.0112	-11.9628	0.0000	-0.1564	-0.1124
video	-0.0703	0.0284	-2.4782	0.0132	-0.1259	-0.0147

```
confusion matrix:
 [[4494  933]
 [1824 4915]]
```

```
Model: Logistic Regression
Accuracy: 0.773384843005
Precision: 0.840458276334
Recall: 0.729336696839
F1-score: 0.780964487169
```

The second trail slightly improved the results and got a higher accuracy.

After that, we found that projects in different industry will have a different fate. From the following figure, we can see that projects in Fashion industry have 67.19% to succeed, while projects in Dance industry only have only 25.46% to succeed. Therefore, we decided to include category dummies in our third trail.



Trail 3: Predictor variables: 'updates', 'levels', 'Average', 'comments', 'FAQ', 'video', ‘category’ (dummy)

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
updates	0.2928	0.0058	50.6326	0.0000	0.2815	0.3042
levels	-0.0073	0.0041	-1.7792	0.0752	-0.0154	0.0007
Average	-0.0011	0.0000	-27.6236	0.0000	-0.0012	-0.0010
comments	0.1619	0.0057	28.3777	0.0000	0.1507	0.1731
FAQ	-0.0958	0.0118	-8.1335	0.0000	-0.1189	-0.0727
video	0.1449	0.0330	4.3972	0.0000	0.0803	0.2095
category_Comics	-1.5440	0.1116	-13.8356	0.0000	-1.7627	-1.3253
category_Dance	0.5803	0.1163	4.9898	0.0000	0.3524	0.8083
category_Design	-1.9538	0.0938	-20.8265	0.0000	-2.1377	-1.7699
category_Fashion	-1.5606	0.1027	-15.1888	0.0000	-1.7620	-1.3592
category_Film & Video	-0.6385	0.0401	-15.9068	0.0000	-0.7171	-0.5598
category_Food	-1.1733	0.0908	-12.9250	0.0000	-1.3512	-0.9953
category_Games	-2.8867	0.1224	-23.5847	0.0000	-3.1266	-2.6468

```
confusion matrix:
 [[4352 1075]
 [1466 5273]]
_________________________________________________
Model: Logistic Regression
Accuracy: 0.791139240506
Precision: 0.830655324512
Recall: 0.782460305683
F1-score: 0.805837854359
_________________________________________________
```

Again, the third trail improved our model and got a higher accuracy can F1-score.

Trail n: Predictor variables: 'goal', 'updates', 'levels', 'Average', 'comments', 'FAQ', 'length(s)', 'duration', ‘category’ (dummy)

We continuously apply different combinations of variables, and gradually improve our accuracy. Finally, we got the best trail when we include this combination into our model. The result is:

```
confusion matrix:
 [[4496  931]
 [1171 5568]]
_________________________________________________
Model: Logistic Regression
Accuracy: 0.827223409502
Precision: 0.856747191876
Recall: 0.826235346491
F1-score: 0.841214684998
```

For complete trails and results, please see the attached python notebook ”Logistic Regression”.

3.2 Decision Tree, Random Forest & Extra Tree

3.2.1 Model Explanation

Model 1: Decision Tree

To further predict the success of Kickstarter projects, we introduce the Decision Tree model, which can be used to predict categorical target variables with sufficient accuracy. By testing a sequence of features, the data is split into groups with the lowest entropy using information gain.

Model 2: Random Forest

Since Decision Tree model may lead to overfitting problem, we further build the Random Forest model, which creates hundreds of trees by taking samples randomly from the train data. Then final prediction is made by taking the majority of the single prediction outcomes, which increases the credibility of prediction.

Model 3: Extra Tree

We want to explore more modeling methods similar to the principle of Decision Tree but with different selection criteria, so we further introduce the Extra Tree. Extra Tree stands for Extremely Randomized Trees. It also uses bootstrap and chooses samples randomly as Random Forest to avoid overfitting. The difference lies in the split of trees. Random Forest is more deterministic where the next split is chosen using information gain, while Extra Tree is extremely random when choosing where to split within the domain.

We want to test which one of the three models is more accurate in predicting our project, which will be shown in the result evaluation.

3.2.2 Parameter Design

The target variable in our project is the funding status (the dummy variable *status_successful* is 1 if the project is successful, 0 if it is failed). To find the most informative predictor variables, we try different combinations.

Trial 1: Predictor variables: ‘goal', 'levels', 'updates', 'comments' ,'duration', 'Average', 'CV', 'date', 'year', 'category'(dummy), 'subcategory'(dummy), 'location'(dummy), 'weekday'(dummy), 'month'(dummy)

Trial 2: Predictor variables: ‘goal', 'levels', 'updates', 'comments' ,'duration', 'Average', 'CV', 'date', 'year', 'category'(dummy), 'subcategory'(dummy), 'location'(dummy), 'weekday'(dummy), 'month'(dummy), 'FAQ', 'video', 'length(s)'

Trial 3: Predictor variables: ‘goal', 'levels', 'updates', 'comments' ,'duration', 'Average', 'CV', 'date', 'category'(dummy), 'month'(dummy), 'FAQ', ‘length(s)'

In Trial 1, we put all variables that we assume to be potentially correlated with the success of the crowdfunding project before crawling. In Trial 2, we add 'FAQ', 'video', 'length(s)' these three variables crawled from the website which we assume may be important predictors. We found that Trial 2 has slightly higher F1 score (0.86) than Trial 1 (0.85), which verifies our assumption. Based on the importance ranking generated in Trial 2, we further exclude some variables that are less important for prediction, which are 'location'(dummy), 'subcategory'(dummy), year, 'weekday'(dummy) and 'video'. The rest of variables which are more informative are used as predictors in Trial 3. This simplified parameter setting generates a shallower tree with higher accuracy than Trial 2. We will discuss the results of Trial 3 in detail with more model testing in later section.

The confusion matrix for Trial 1 and 2 and the top variables with high importance are presented in Appendix (Figure 14 - Figure 16).

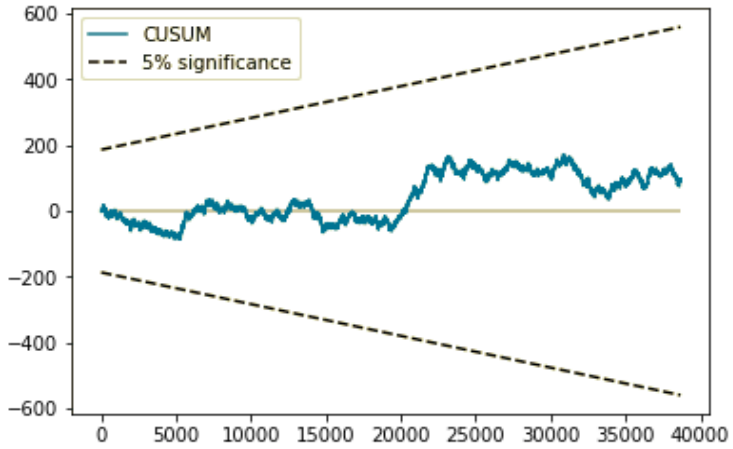
3.3 Linear Regression

To further find out how to raise more money, linear regression model is adopted here as we take funded_percent (pledged money / goal), which is a continuous parameter, as the dependent variable. Because the target variable has some extreme point as stated in EDA, we try the before scaling and after scaling model to improve the performance. Since for linear regression, the selection of independent variable will largely determine the performance of the model. Our general method to try different set of independent variable, try to find the largest adjusted R-squared and, at the meantime, try different model in statsmodels to make stable coefficient.

Method1: By using OLS method in statsmodels, numerous set of independent variables including dummy variables creating based on category, location and etc, we finally find the optimal model with best adjusted R-squared. The independent variables are posted below.

	Final Model	Trial 1	Trial 2	Before Scaling
const	5.0098***	4.9843***	4.9888***	5.0521***
updates	0.0241***	0.0235***	0.0237***	0.0286***
Average	-0.0001***	-0.0001***		-0.0001***
length(s)	8.155E-05***	0.0001***		2.538E-05 (p-value = 0.104)
category_Comics	-0.1051***		-0.0914***	-0.0569***
category_Dance	0.1322***		0.1419***	0.1077***
category_Design	-0.122***		-0.1195***	0.0195 (p-value = 0.119)
category_Fashion	-0.1467***		-0.1489***	-0.1495***
category_Film & Video	-0.0397***		-0.0612***	-0.0581***
category_Food	-0.0486***		-0.0585***	-0.0594***
category_Games	-0.1976***		-0.1993***	-0.102***
category_Music	0.071***		0.0703***	0.059***
category_Photography	-0.084***		-0.0835***	-0.1009***
category_Publishing	-0.1317***		-0.1287***	-0.1394***
category_Technology	-0.151***		-0.165***	-0.0686***
category_Theater	0.1053***		0.108***	0.0848***
Adjusted R-squared	0.242	0.188	0.222	0.202
Stability (recursive LS)	Stable	Stable	Stable	Stable

Method 2: Using recursive LS method, we can check the stability of parameter coefficient. The result can be shown in CUSUM statistic. In the plot below, the CUSUM statistic does not move outside of the 5% significance bands, so we fail to reject the null hypothesis of stable parameters at the 5% level.



4. Results & Evaluation:

4.1 Logistic Regression Results

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
goal	-0.0002	0.0000	-43.4194	0.0000	-0.0002	-0.0002
updates	0.3352	0.0065	51.3031	0.0000	0.3224	0.3480
levels	0.0391	0.0047	8.3221	0.0000	0.0299	0.0483
Average	0.0001	0.0000	2.9879	0.0028	0.0000	0.0002
comments	0.2590	0.0071	36.7173	0.0000	0.2452	0.2728
FAQ	-0.0666	0.0132	-5.0416	0.0000	-0.0925	-0.0407
length(s)	0.0004	0.0001	2.7861	0.0053	0.0001	0.0006
duration	-0.0182	0.0009	-21.3831	0.0000	-0.0199	-0.0166
category_Comics	-0.9989	0.1257	-7.9450	0.0000	-1.2454	-0.7525
category_Dance	1.1916	0.1241	9.6007	0.0000	0.9484	1.4349
category_Design	-1.0048	0.1060	-9.4832	0.0000	-1.2125	-0.7971
category_Fashion	-0.9226	0.1118	-8.2533	0.0000	-1.1417	-0.7035
category_Film & Video	0.2257	0.0472	4.7805	0.0000	0.1332	0.3182
category_Food	-0.1662	0.1065	-1.5608	0.1186	-0.3750	0.0425
category_Games	-2.2908	0.1423	-16.0965	0.0000	-2.5698	-2.0119
category_Music	0.5175	0.0500	10.3549	0.0000	0.4196	0.6155
category_Photography	-0.4800	0.0946	-5.0749	0.0000	-0.6654	-0.2946
category_Publishing	-0.6134	0.0628	-9.7603	0.0000	-0.7366	-0.4902
category_Technology	-1.3128	0.1798	-7.3026	0.0000	-1.6652	-0.9605
category_Theater	1.1363	0.0741	15.3434	0.0000	0.9911	1.2814

```
confusion matrix:
 [[4496  931]
 [1171 5568]]
```

```
Model: Logistic Regression
Accuracy: 0.827223409502
Precision: 0.856747191876
Recall: 0.826235346491
F1-score: 0.841214684998
```

For the model that provides us with the highest accuracy, we found that nearly all the predictors are significant, except dummy variable [‘*category_food’*]. Moreover, in this model, [‘goal’, ‘FAQ’, ‘duration’] and some dummies like [category_comics, category_design, category_fashion, category_games] are negatively related to the final status, which means, if a project is in this industry, with high goals, more FAQs, and long duration, it’s highly probable to fail. On the other hand, if a project updates frequently, and are commented frequently, and has video to promote, it will have more probability to succeed.

4.2 Decision Tree, Random Forest & Extra Tree Results

After comparison, we use Method 3 as our predictor variables. We run Decision Tree model, Random Forest model and Extra Tree model separately with parameters in Method 3 using training dataset (randomly choose 70% of the entire dataset) and test the model using test dataset (the rest 30% of the dataset).

```
confusion matrix:
 [[4118 1354]
 [1446 5248]]
```

```
Model: Decision Tree
Accuracy: 0.7698504027617952
Precision: 0.7949106331414723
Recall: 0.7839856587989245
F1-score: 0.789410348977136
```

Decision Tree Model

```
confusion matrix:
 [[4419 1053]
 [ 831 5863]]
```

```
Model: Random Forest
Accuracy: 0.8451421995725793
Precision: 0.8477443609022557
Recall: 0.8758589781894234
F1-score: 0.8615723732549596
```

Random Forest Model

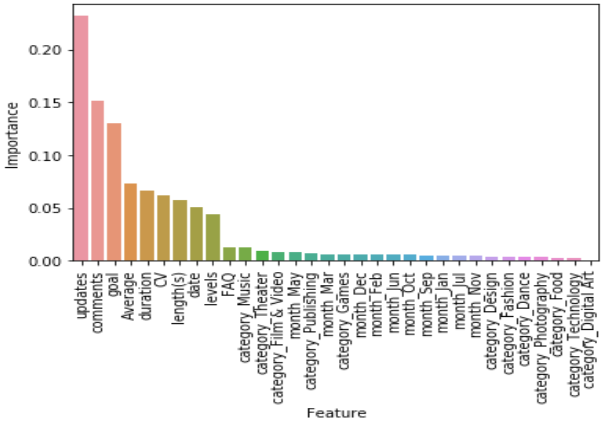
```
confusion matrix:
 [[3591 1881]
 [1875 4819]]
```

```
Model: Extra Tree
Accuracy: 0.6912707545618938
Precision: 0.7192537313432836
Recall: 0.7198984164923813
F1-score: 0.7195759295206808
```

Extra Tree Model

Comparing the test results of these three models, we found Random Forest has the highest Accuracy and F1 score, then Decision Tree comes after, Extra Tree being the lowest. To interpret, Decision Tree model lacks randomness compared with Random Forest, so the result may be rather biased. As for Extra Tree model, it may not work well if there is a high number of noisy features. In our case, the dataset has high dimensions with dummy variables. Extra Tree may also reduce variance because of further decorrelation of the trees.

The result of Random Forest shows 4419 True Positive predictions, 1053 False Positive predictions, 831 False Negative predictions and 5863 True Negative predictions, which refers to 85% of total correct predictions, 85% of positive correct predictions, 88% correct predictions of successful projects and overall F1 score to be 0.86.



We further plot the bar chart of features based on importance level in Random Forest model. It shows that 'updates', 'comments', 'goal' has significant higher importance in distinguishing successful projects from failed ones. Dummy variables 'category' and 'month' has less importance, but among them, Music, Theater and Film & Video categories for 'category' and May and March for 'month' have better performance.

This result indicates essential information for start-up initiators. First, they should attach great importance to information transparency and timeliness by updating their websites with news and latest progress. Second, they should put more effort in maintaining interaction and communication level by replying comments and FAQs timely or even creating comments by themselves. Third, they must have rational expectation of the result by setting proper goals. If the goal is set too high, they will face higher probability that the crowdfunding will fail in which situation all money pledged will be refund to the backers. However, if the goal is set too low on purpose, for example $1, as we see in some projects, chances are that you easily succeed, so you have to give the corresponding reward as promised, even if the funding is not enough to make the reward product or compensate for the cost of the reward without economics of scale.

4.3 Linear Regression Results

Although the final model is optimal in terms of Adjusted R-squared (0.242). However, it is still a weak model since it just explains 24.2% of variations in target variables. Moreover, the conditional number is large (8.82e+03), which suggests that there are strong multicollinearity or other numerical problems. To further dig out the pattern for each parameter, partial regression plot is cited in the appendix), which also suggests some weakness in this model.

4.4 Evaluation

In total, we applied three types of models to analyze this problem.

Logistic regression was used to predict the results, and also to reveal the relationship of predictors and target variable. Compared to logistic regression, random forest model (ensemble of decision tree model) provided us with a higher accuracy and F1-socre, which means random forest has a better ability to predict. However, although Random Forest showed us the importance rank of predictors, it didn't show us the

positive and negative relationship between each predictor and target variable, as what logistic regression showed to us.
As for the linear regression, although it overcame the drawback of logistic regression and random forest (the target variable for logistic regression and random forest is binominal and categorical, cannot tell us how to increase the money raised), linear regression didn't provide us a good prediction. It still a weak model.

5. Recommendations:

Based on the result from logistic regression, random forest and linear regression, recommendations to the three business problems can be concluded here.

5.1 Interaction Matters:

As the significantly positive coefficient of FAQ, updates, comments suggests, the interaction between fund raisers and backers will be of vital importance to make your project successfully funded. Therefore, we will recommend project creators to make initiatives to improve the transparency and smooth communication channels between creators and backers. In this way, backers will see project is on its way to success, the feedbacks from them are valued and creators are authentic to land the idea down to earth.
Some sample initiatives including:

- Monthly Updating project progress
- Draw a dynamic timeline of the project, weekly updates.
- Daily check FAQ and actively react to the feedback
- Create backer comunity and create facebook and twitter to communicate with them

Lessons from successful case:
1.FACE 2 FACE a documentary film, pledged $77,547 with goal $50,000, updates 155
https://www.kickstarter.com/projects/KatherineBrooks/face-2-facebook-a-documentary-film
2.Pebble: E-Paper Watch for iPhone and Android, pledged $10,266,845 with goal $100,000, comments 15,660
https://www.kickstarter.com/projects/getpebble/pebble-e-paper-watch-for-iphone-and-android
3.Butcher Shop at Sugar Mountain Farm - Pastured Pigs, pledged $33,456 with goal $25,000, FAQ 39
https://www.kickstarter.com/projects/sugarmtnfarm/building-a-butcher-shop-on-sugarmountainfarm

5.2 Awarding in levels:

Other than interaction, the level of awards can essentially improve the success chance of the project. As shown in the dataset, project creators award backers in different ways. Many successful project awards in six or seven levels. More levels will keep an option open for backers-- investing a little more, get awards a little more. This positive reaction will always promote more investment and thus rise the chance to successfully raise money. Therefore, our logistic regression model result strongly recommend creators adopt level awarding strategy and delicately design their awards.

Some sample initiatives including:

- Award different product with different functions in different levels
- Award product in different providing date
- Award some spiritual feedback including design books or team greeting gifts
- Adding random gifts to higher level

Lessons from successful case:
1. Speakeasy Dollhouse, $16,811 pledged with $10,000 goal, 10 levels
https://www.kickstarter.com/projects/183801348/speakeasy-dollhouse
2.God Help The Girl - Musical Film, $121,084 pledged with $100,000 goal, 6 levels

https://www.kickstarter.com/projects/godhelpthegirl/god-help-the-girl-musical-film

5.3 Using our random forest to make a reasonable funding goal:
By inputting your project current presentation features, our random forest can output the current stage of your project, whether or not your project will be successfully funded. Moreover, if output is unsuccessful, by looking at decision trees, suggestion will be given based on why your branch is more likely to be unsuccessful. For example, maybe the average awarding is a bit low. Maybe the category of your project suggests you to adjust your goal lower.

5.4 Video and length of it will boost your success:
Finally, based on three models, video will increase your funded percentage and also your success possibility. Therefore, creators are highly recommended to delicately design their video and make it as informative as possible. In this way, backers will enjoy your video and at the same time, have more probability to be attracted by your idea.

Some initiatives including:
- If not too long, the longer and effort you put on the video, more possible you will pledge more money
- Tell story instead of plainly talking

Lessons from successful cases:
1.JourneyQuest Season 2, $113,028 pledged of $60,000 goal, video length: 9:41
https://www.kickstarter.com/projects/zombieorpheus/journeyquest-season-2
2.Be a Part of Dan Macaulay's New CD!, $13,158 pledged of $10,000 goal, video length: 6:20
https://www.kickstarter.com/projects/1702996164/be-a-part-of-dan-macaulays-new-worship-album

6. Limitations and Further Research:

6.1 Limitations:
For linear regression and logistic regression, the biggest drawback that we fail to overcome, is the multi-collinearity among all the predictors. Due to the limitation of time and technology, we didn't finish the application of stepwise regression and ridge regression, which are outstanding in selecting variables. Therefore, for our further research, we will also apply different methods to minimize the multi-collinearity, and build a better explanatory model.

6.2 Further Research:
Our group will further move into text mining. The major motivation to adopt text mining is that the project description in the website is the most informative data. Therefore, we want to use text mining, which will transform these string data into more informative data and then improve our model performance, since there are more information added into the model.

Moreover, we will collect more data in terms of the rewards, including the form of awards (product, money, or books), the total value of each rewards. This is primarily because the awards are also one of the most important initiatives for backers to invest in Kickstarter platform. A further study on this will give more specific recommendation on the awarding strategy.

Finally, if possible, we will adopt more recent data to renew our random forest, logistic regression and linear regression model. This is motivated by the dynamic changes happening in Kickstarter platform, which suggests more recent data will have more predictive power.

Appendix

Figure 1. Countplot: Distribution of Categories

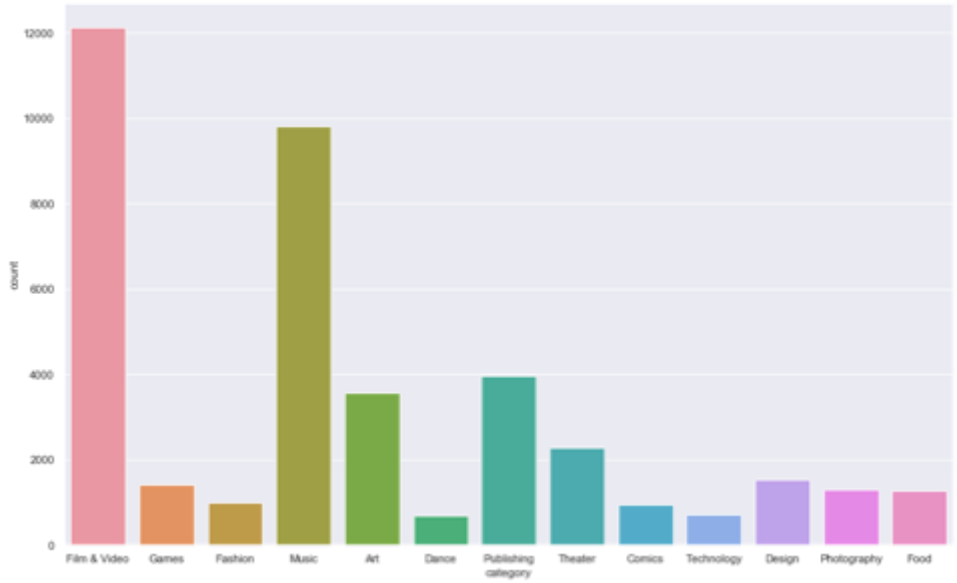


Figure 2. Bubble Plot: Total Fund Raised by Category

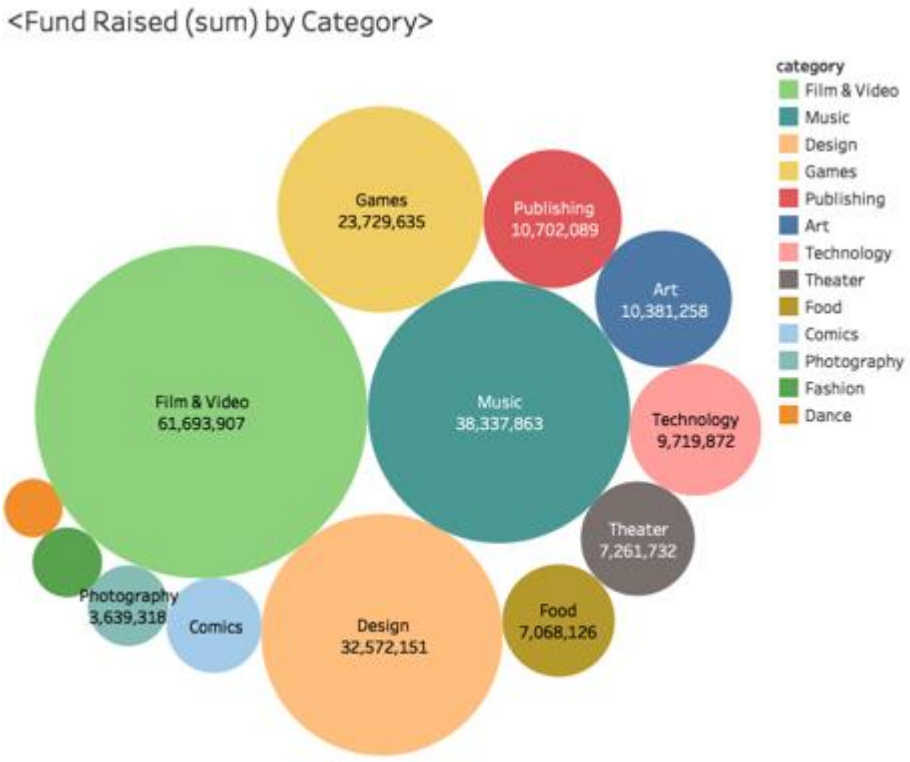


Figure 3. Bar Chart: Averge Fund Raised Per Backer by Category

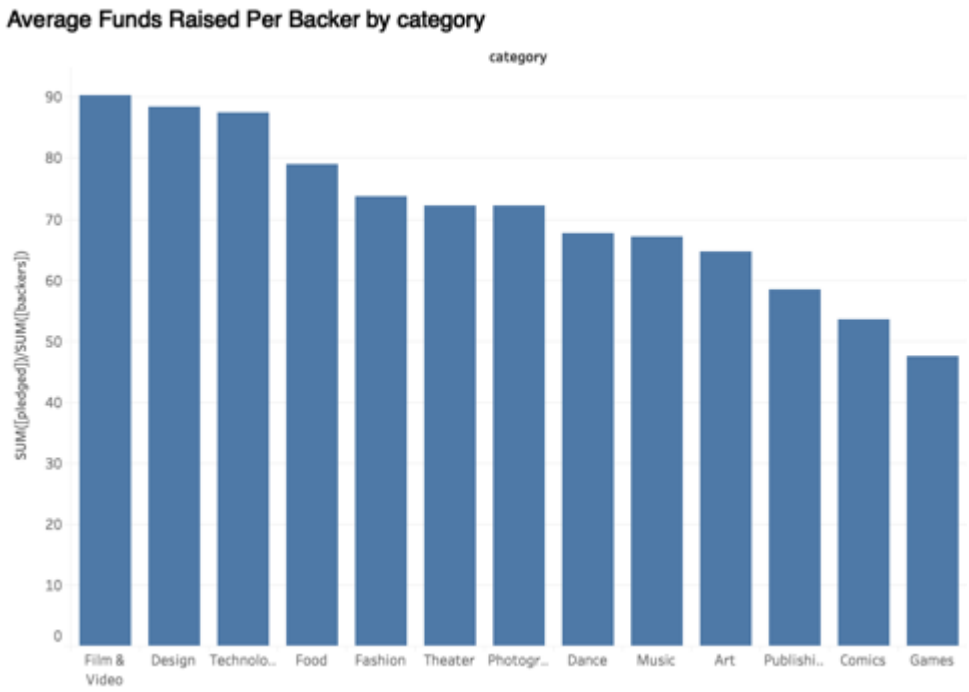


Figure 4. Bar Chart: Success Rate by Category

(Then we calculate the success rate of each category and found that though Film & Video received the highest fund, Dance, Theater and Music are the categories have the highest success rate indicating entertainment projects are in general easier to succeed.)

Figure 5. Map: Project Initiator Location Worldwide

Figure 6. Map: Project Initiator Location Within the US

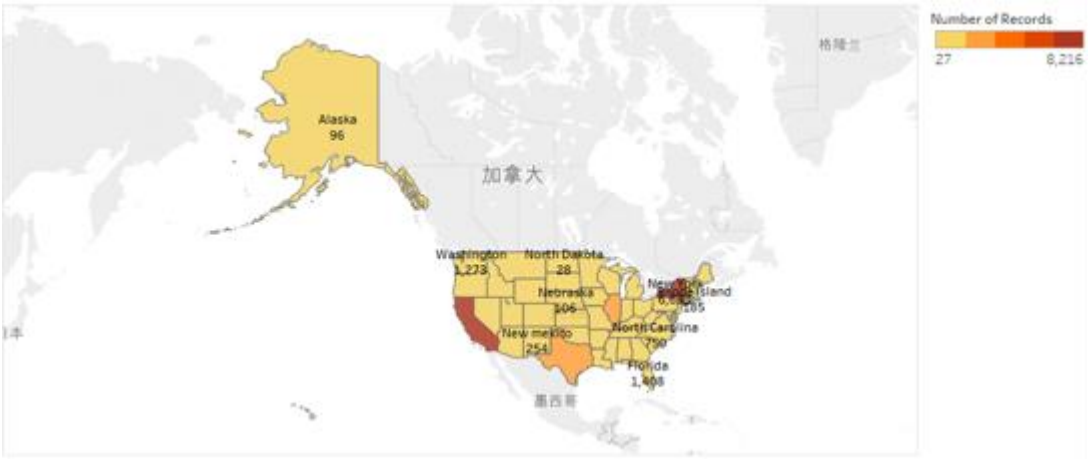


Figure 7. Bar Chart: Top 13 Locations

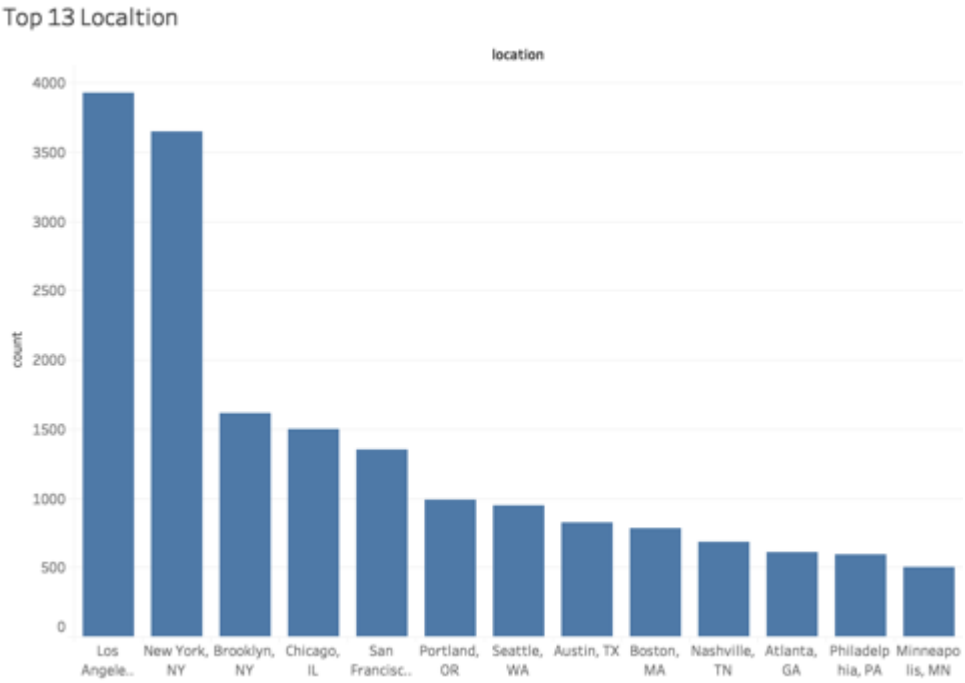


Figure 8. Bar Chart: Distribution of Projects Year-Wise (by Category)

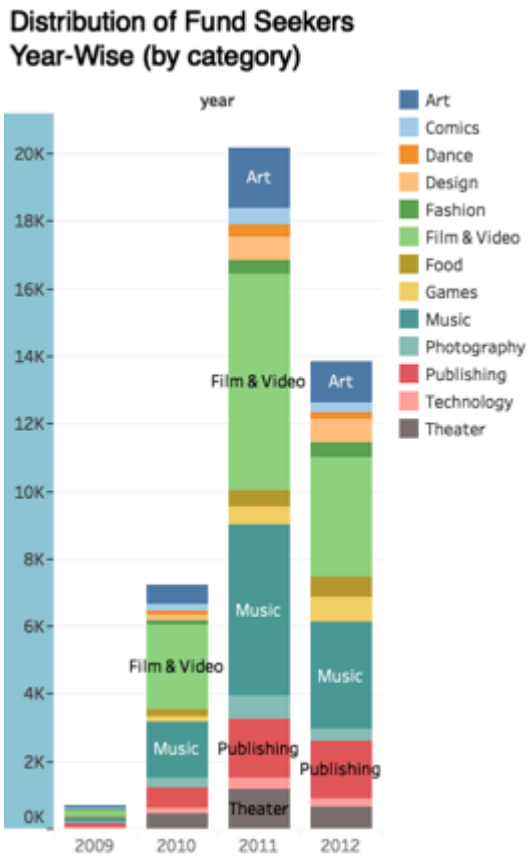


Figure 9. Bar Chart: Top 12 Fund Raisers

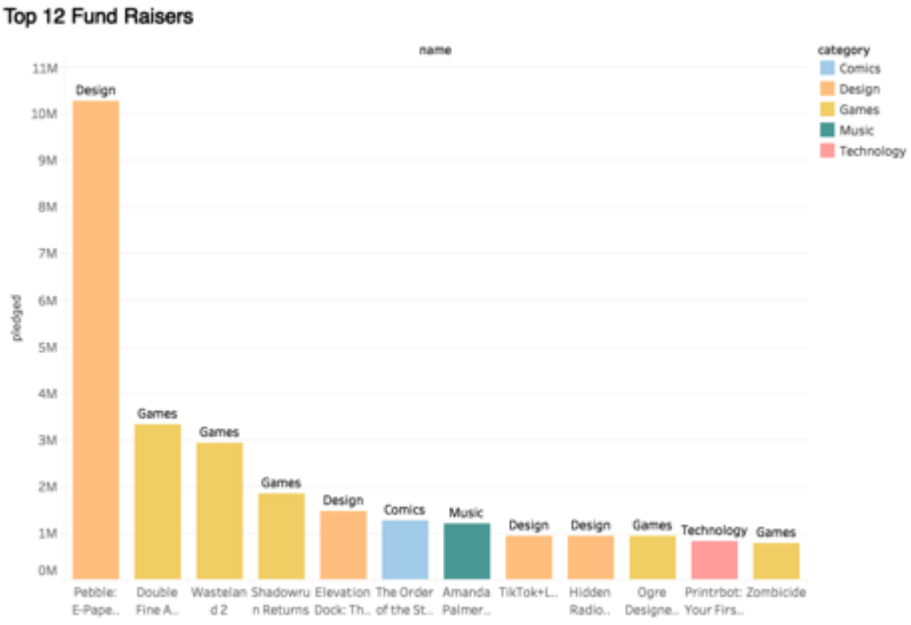
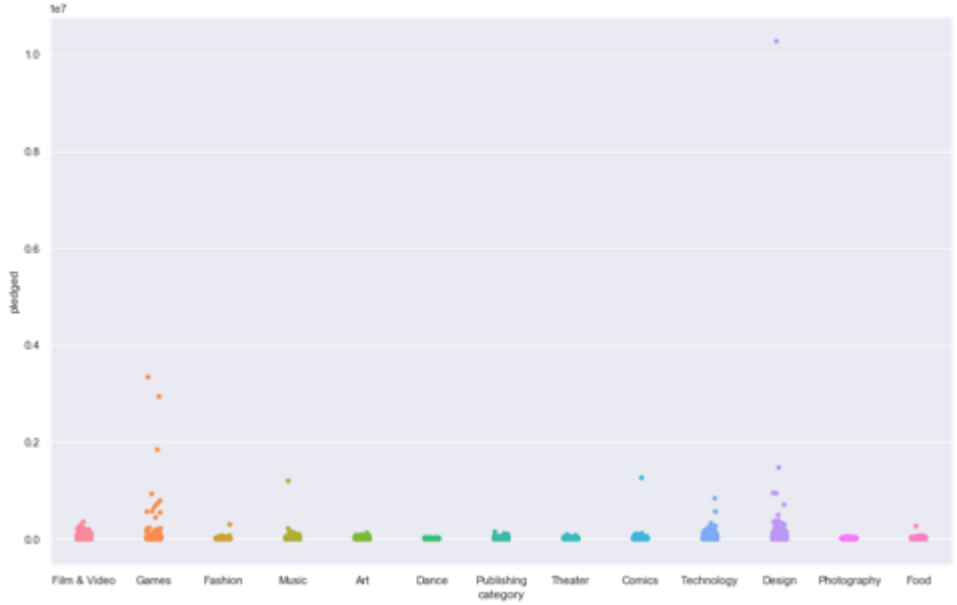


Figure 10. Stripplot: Fund Pledged by Category



(The stripplot showing category-wise money pledged, we found there are actually some outliers. Some Design projects have significantly higher money pledged. The highest one is an E-Paper project raising more than $10 million, while the great majority of projects receive funds in hundreds or thousands.)

Figure 11. Boxplot: Fund Pledged by Category

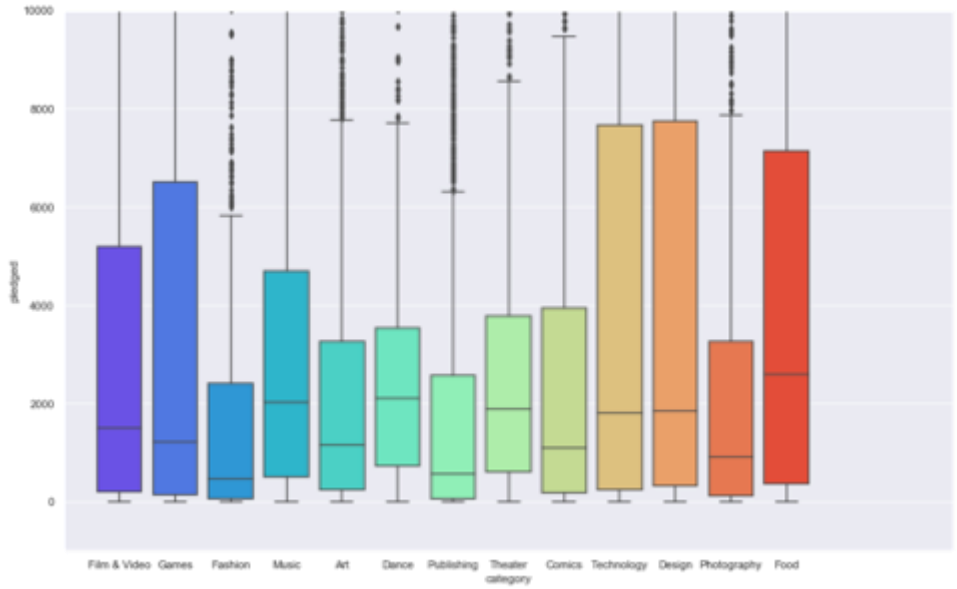


Figure 12. Stripplot: Fundingt Duration by Status

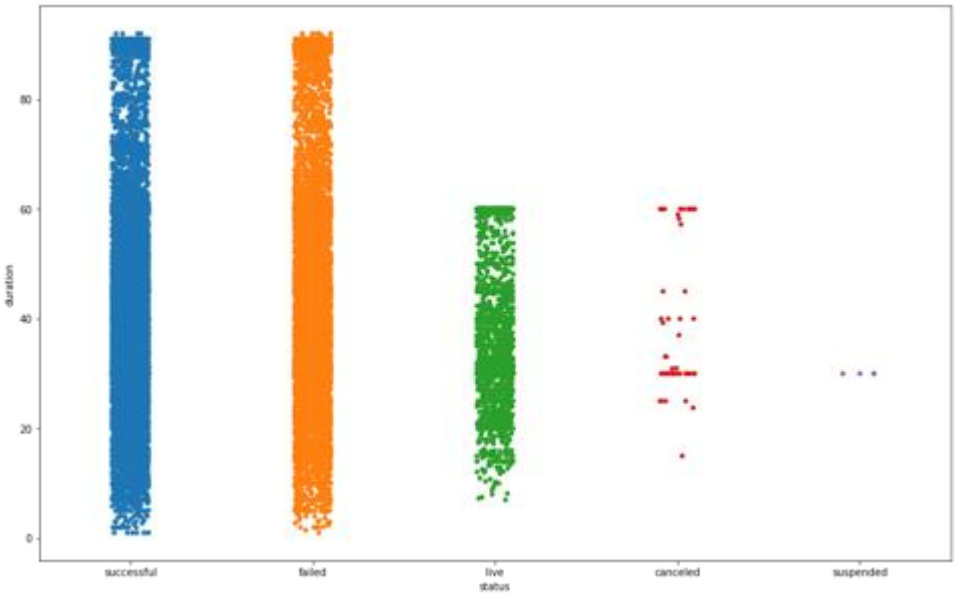


Figure 12. Boxplot: Fundingt Duration by Status

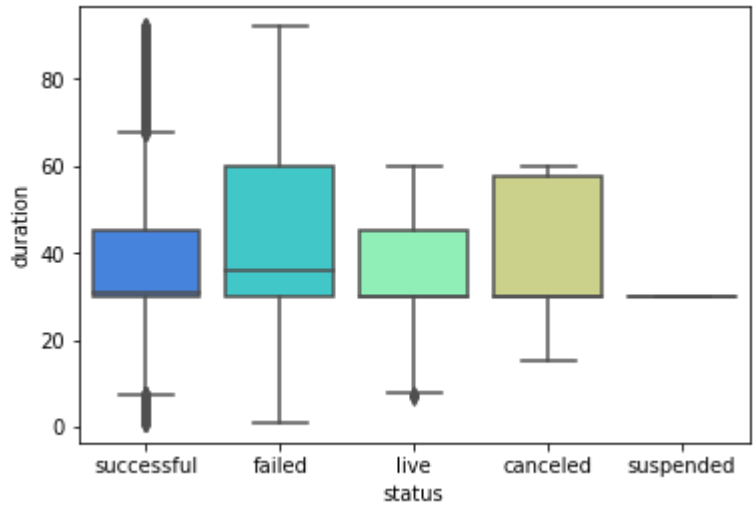


Figure 13. Scatterplot: Relationship Between Backers and Goal

Figure 14. Confusion Matrix of Trial 1 Using Random Forest

```
confusion matrix:
 [[4439 1133]
 [ 894 6096]]
_____________________________________________
Model: Random Forest
Accuracy: 0.8386403438942843
Precision: 0.8432701618481118
Recall: 0.8721030042918455
F1-score: 0.8574442647162247
_____________________________________________
```

Figure 15. Confusion Matrix of Trial 2 Using Random Forest

```
confusion matrix:
 [[4315 1157]
 [ 737 5957]]
_____________________________________________
Model: Random Forest
Accuracy: 0.8443202367253
Precision: 0.8373629463030644
Recall: 0.8899014042426053
F1-score: 0.8628331402085748
_____________________________________________
```

Figure 16. Top 15 Feaures With Highest Importance in Trial 2

```
[Text(0,0,'updates'),
 Text(0,0,'goal'),
 Text(0,0,'comments'),
 Text(0,0,'Average'),
 Text(0,0,'duration'),
 Text(0,0,'CV'),
 Text(0,0,'length(s)'),
 Text(0,0,'date'),
 Text(0,0,'levels'),
 Text(0,0,'year'),
 Text(0,0,'category_Theater'),
 Text(0,0,'category_Music'),
 Text(0,0,'category_Games'),
 Text(0,0,'location_New York, NY'),
 Text(0,0,'FAQ'),
```

Figure 17: Partial Regression Plot: