Analysis on Crawled Kickstarter Data

Tingjun Zhang (tz976) Chang Liu (cl3869)

Abstract -- In this project, we did data summary on crawled data from Kickstarter, a fundraising website that contains projects informations by using big data techniques such as Spark for data analyzing. The results are expected to show some relationships between project state and factors such as category, geo-distribution, goals, etc. We also use other machine learning tools such as sklearn and pytorch for predictions on the possibility that if a project will success at the startpoint.

Keywords: Kickstarter, analysis, predictions, Spark, Machine Learning

Table of contents

Analysis on Crawled Kickstarter Data **1**

I. INTRODUCTION 4

II. DATA SUMMARY 4

2.1 Frequency of project states 4

2.2 Average amount of pledged money for different states of project 5

2.3 Category analysis 6

2.3.1 Category frequency 6

2.3.2 Category success amount 7

2.3.3 Category failure rate 7

2.3.4 Subcategory frequency 8

2.4 Geo distribution of projects 10

2.5 Average backers of different states of project 12

2.6 Average number of comments for different states of project 13

III. PREDICTIONS 13

3.1 Prediction on the success of a project 13

3.1.1 Basic prediction 15

3.1.2 Importance of different features 15

3.1.3 Prediction Update 15

3.1.3 Conclusion 16

3.2 Prediction on the amount of money pledged eventually 17

3.2.1 training set 1 18

3.2.2 training set 2 19

3.2.3 Conclusion 21

IV. APPENDIX 21

4.1 Data source 21

4.2 Web crawler 22

4.2.1 Project urls 22

4.2.2 Crawling 22

4.2.3 Data storing 22

4.3 Package usage overview 23

4.4 Data summary 23

4.5 Predictions 23

I. INTRODUCTION

Kickstarter is a famous crowdfunding website in United States, many people come here to start their own projects. Some projects achieve great success while some projects can hardly get any attention. The reasons are various, but there seems to be some relationships between the final results and some features of those projects. It would be nice if we can figure them out and provide proper guidance based on them for both project creators and backers.

The goals of this project are doing data summary and analysis based on crawled data from kickstarters, and predicting whether a live project will success or fail eventually, and how much money will be expected to be pledged in the end. We retrieve data from Web Robot and we write our own customized crawler to extend the features that we need. Based on the crawled data, we found some interesting relationships between the features and the success rate of the projects.

II. DATA SUMMARY

Our dataset includes details of 203338 projects in total. The time periods of those projects are mostly between 2014 and the present. We did some exploratory data analysis on the dataset to get a better understanding and briefly summarize relationships among each feature using Spark. Below is an overview of our results.

2.1 Frequency of project states

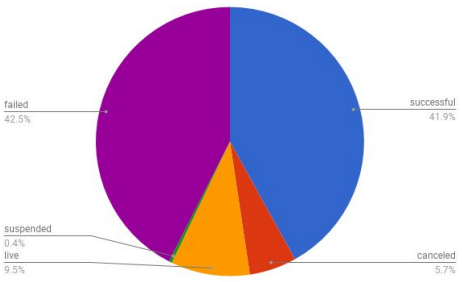

Fig 2.1 State frequency(1)

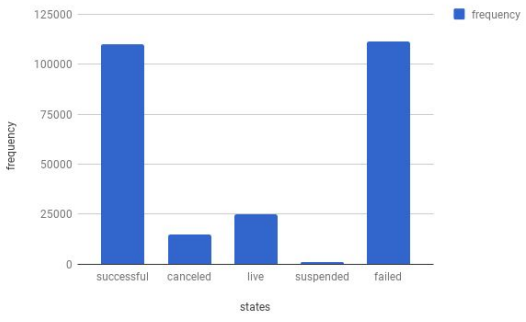

Fig 2.2 State frequency(2)

Generally, the number of successful and failed projects are almost the same. In our results, failed projects (42.5%) is slightly more than successful projects (41.9%). 5.7% projects are canceled and 0.4% are suspended. There are about 19318 live projects at present.

2.2 Average amount of pledged money for different states of project

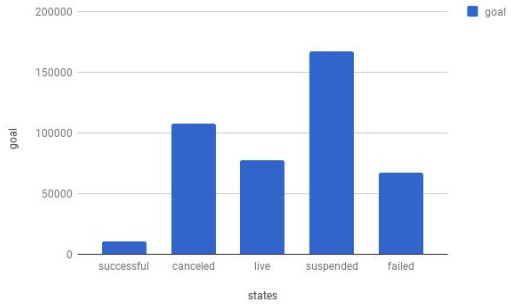

Fig 2.3 Goals

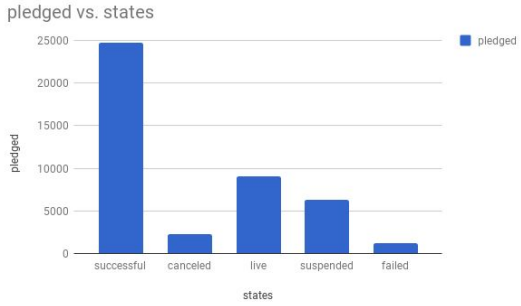

Fig 2.4 Pledged money

The average amount of goal for successful projects is around $10,000. Other kinds of projects (failed, canceled, suspended) tend to have larger amount of goal compared with successful projects. Failed projects have a goal about 6 times more than successful projects. It shows that projects with smaller amount of goal are more likely to achieve success.

The amount of pledged money of successful projects is about $25,000 which is 2.5 times more than their average amount of goals. Based on this result, it is probably reasonable to suggest project starters to set a smaller amount goal than their expected amount.

2.3 Category analysis

2.3.1 Category frequency

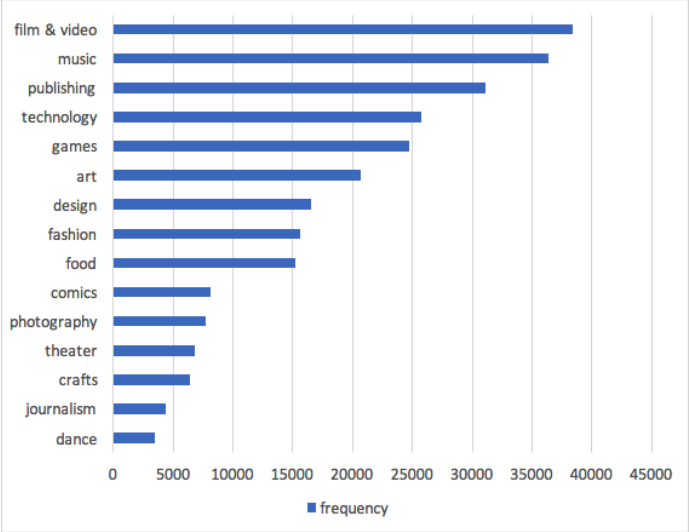

Fig 2.5 Category frequency

The statistics show that film & video, music and publishing are the most popular categories with many related projects, while dance and journalism are in the minority.

2.3.2 Category success amount

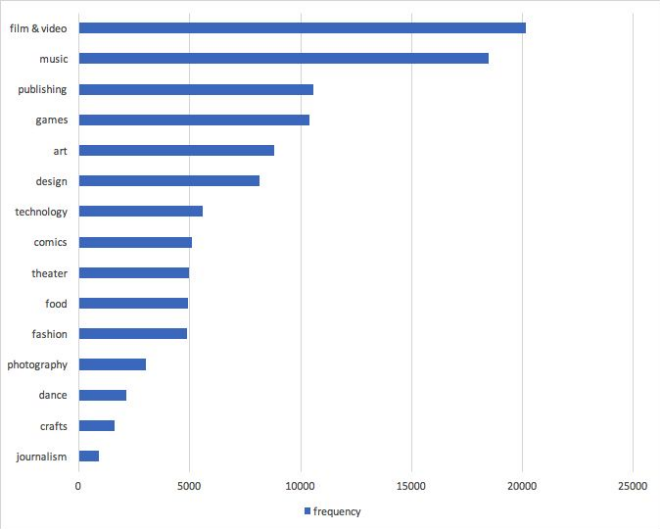

Fig 2.6 Category success statistics

From this chart, we can find that the ranking is very similar to the previous one. It is noteworthy that the ranking of technology has an obvious drop in this chart. For some reason, though there are many projects launched in technology category, the success rate is not as good as games and art category.

2.3.3 Category failure rate

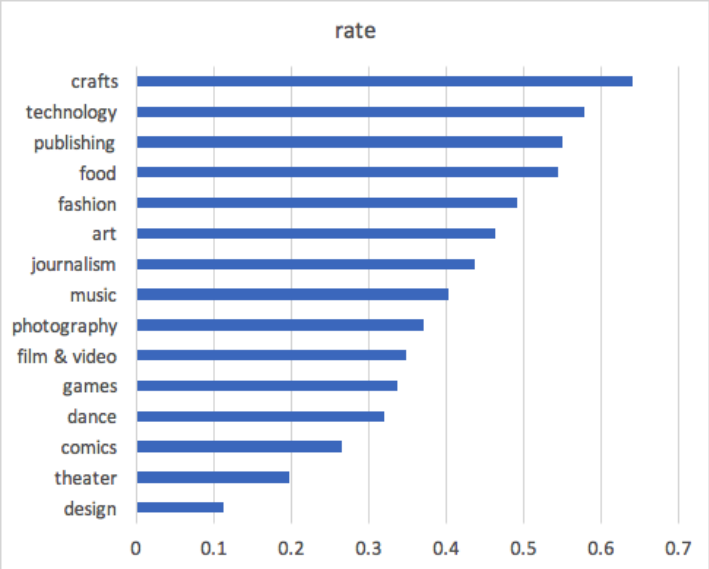

Fig 2.7 Category failure rate chart

The failure rate confirms our observation above. Crafts, technology and publishing have the highest failure rate. Although there are many fancy technology projects on kickstarters, most of them are not accomplished. Backers should be aware of the risk before they pledge money for project with the tag of crafts, technology and publishing.

2.3.4 Subcategory frequency

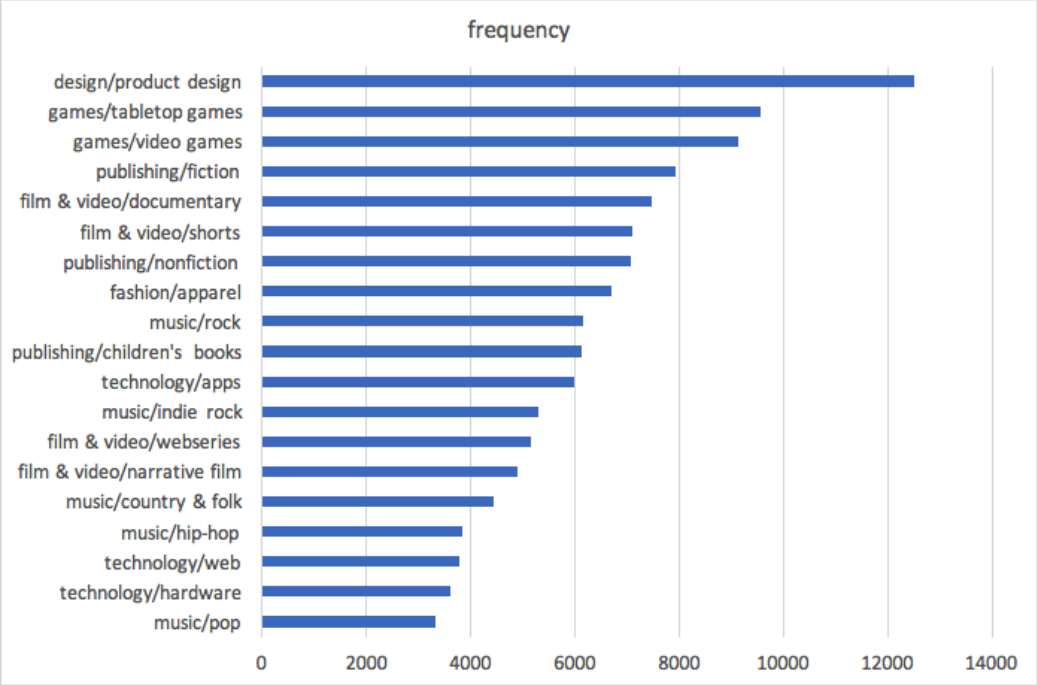

Fig 2.8 Subcategory frequency (top 20 shown in chart)

We are very interested in which sub categories are most popular on kickstarters, so we made statistics for sub categories.
Surprisingly, the most frequent sub category does not belong to those top categories, it is product design. After taking a look at the dataset, we find that category ‘design’ contains 16526 projects and ‘product design’ contains 12496 projects. So most of the projects in design category belong to ‘product design’ which makes it easy for ‘product design’ to take the lead in this ranking.
Besides the product design, we can find the ranking is basically corresponding to the category ranking we have got above. Film & video, games and publishing categories take the majority in this chart.

subcategory	success rate
technology/apps	0.060869565
technology/web	0.063869095
games/mobile games	0.080159181
journalism/video	0.116389549
technology/ software	0.121252947
food/food trucks	0.12311266
crafts/candles	0.126794258
film & video/ action	0.140056022
journalism/web	0.142160845
crafts/ embroidery	0.142857143

subcategory	rate
film & video/ shorts	0.904936015
music/chiptune	0.771428571
film & video/ documentary	0.7557579
games/tabletop games	0.694188963
comics/ anthologies	0.674242424
music/indie rock	0.670690306
dance/ residencies	0.642857143
publishing/ letterpress	0.622222222
music/classical music	0.619451949

Fig 2.9 Worst rate ranking (left) and best rate ranking(right)

From the success rate we calculated from the dataset, we believe certain kinds of subcategory are more likely to succeed compared with others. If you are a project starter, you may want to consider some art related projects, because projects with subcategories under ‘film & video’, ‘music’ and ‘dance’ are more likely to succeed. Technology related projects, on the other hand, don’t have a good performance in success rate ranking. But it is also reasonable, since developing apps, web or mobile game is much more harder than what the creators expected before they actually do it.

The success rate of categories will be considered as an important factor in the prediction part.

2.4 Geo distribution of projects

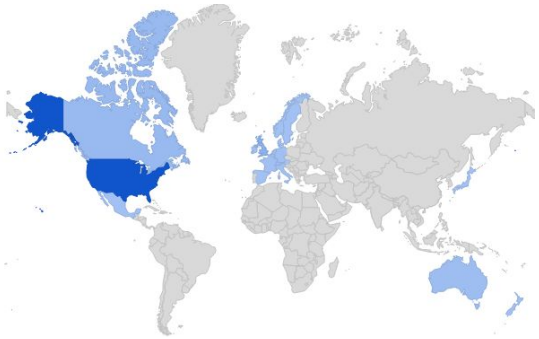

Fig 2.10 Worldwide project distribution map

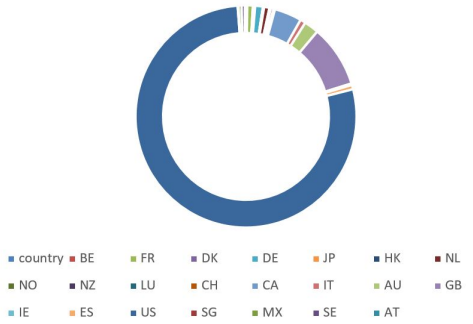

Fig 2.11 Country pie chart

As we could see from the chart above, the projects that launched in Kickstarter are across the globe. The majority of projects are located in the US, and there are also projects located in Europe, Asia, Russia, Australia and New Zealand etc. Also, as over ¾ of projects are located in the US, we mainly focus on projects that are within the US.

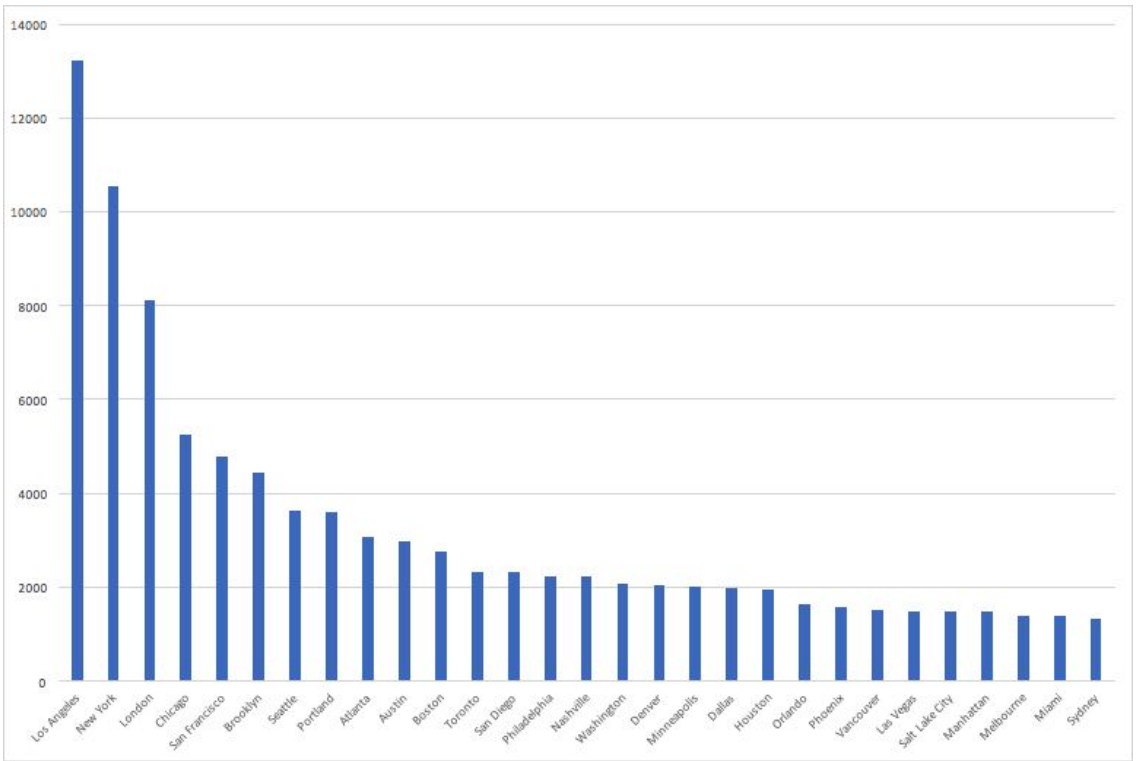

Fig 2.12 Project frequency of main cities

The chart above shows the distribution of projects among the cities in the world, in decreasing sequence. As we can see, the top cities that projects launch at are within the United states. It may due to that kickstarter is a US company, and might have a high popularity among the US but less heard of outside the US. Also, we find that among the top cities, big cities are likely to launch more projects. Such as Los Angeles, NY, Chicago as well as San Francisco. Those are the biggest cities in the US. We assume that there are more opportunities there, thus there are more projects.

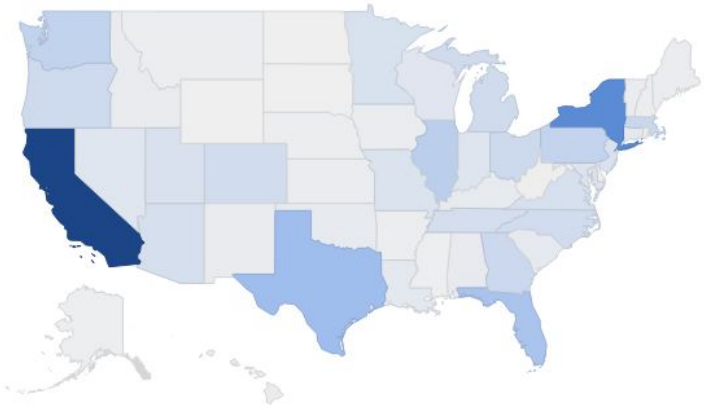

Fig 2.13 Distribution of projects in United States

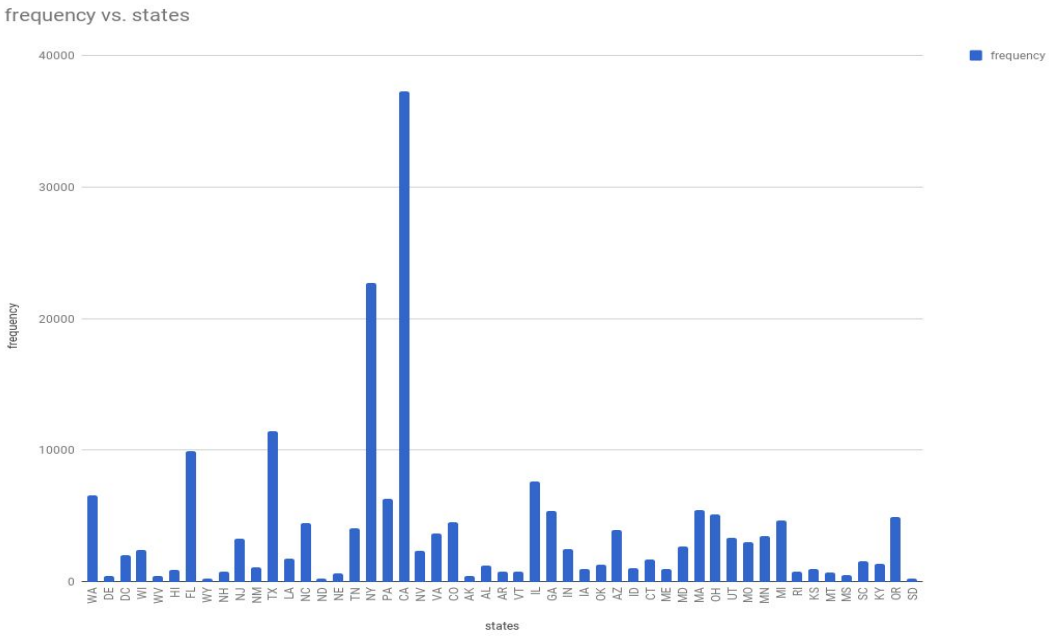

Fig 2.14 Project distribution among states

As we could see from the geo-distributed graph of the projects in United States, both western coast and eastern coast are dense areas where the projects are located. The southern part of the US also has projects, but is much less than the coastal area. In the central of US, there seems a lot less projects, which we assume that the main pillar industry there is not IT or other Internet related business. Also, there might be other reasons such as less population density and geographical customs. Maybe compared with starting a local project online, they would prefer funding offline.

As is shown from the diagram, the top four states with the most number of projects are: CA, NY, TX, and FL. We deduce that certain relationships between the population or other factors have something to do with this distribution. The more population the state has, the more projects are launched in the state. CA has the most amount of projects among all states. We assume that it is due to the reason that it is the birthplace of technology, thus more ideas are implemented here than other places. As for NY, since it is one of the most populated states of United States, there are more markets as well as opportunities here, thus there are a lot of people start their ideas here.

2.5 Average backers of different states of project

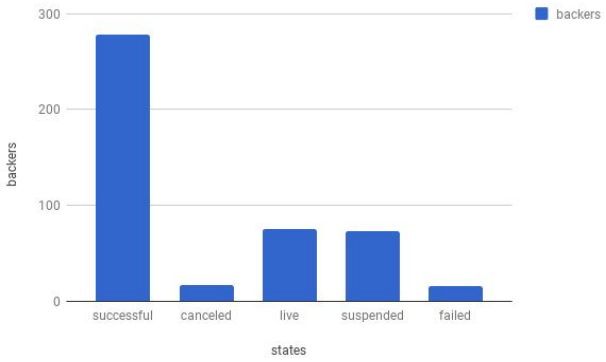

Fig 2.15 Backers count for projects

As is shown in the diagram, there is a strong connection between number of backers of a project and its final state. The average number of backers of a success project is over 250 people, while for failed projects the number is below 20. This is also a strong indicator to predict whether a project will succeed or not.

2.6 Average number of comments for different states of project

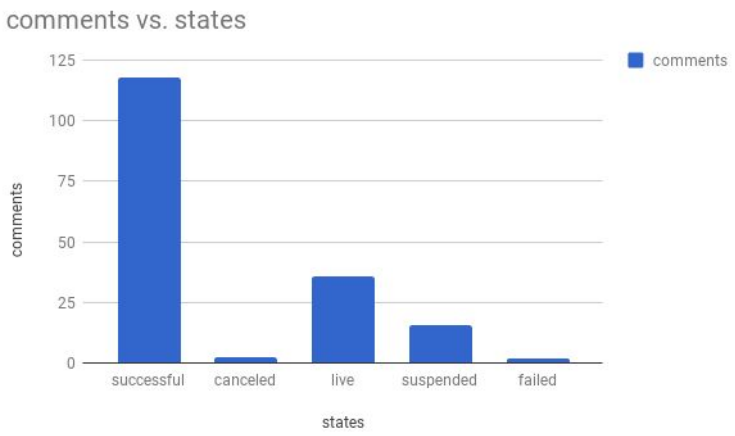

Fig 2.16 Comment num for projects

Previously we assume that there is some correlations between the number of comments with the popularity of the project. We believe the more interactions with others a project starter has, the more likely a project is to get attention and in the end, get more pledged money. It seems that the average comment number did provide facts for that. For a successfully pledged project, it is more inclined to get more attention. The average number of comments is around 110. While on the other hand, failed projects has very few comments, which is around 13. This is a strong feature that could be used to predict whether a project will eventually success or not.

2.7 Creator performance

We assume that the creators themselves have a big impact on the final state of their projects, so we analyze the dataset to get total project number and successful number for each creator. We also calculate the success rate for each creator. Below is the top 20 creators with best performance within Kickstarters.

creator	total	success	rate
167427101	46	46	1
1323060510	30	30	1
1287277253	30	30	1
coolminiornot	26	26	1
jeffdee	21	21	1
1016374822	42	41	0.976744186
1655558466	32	31	0.96969697
578114575	32	31	0.96969697
563681582	28	27	0.965517241
laporte	27	26	0.964285714
cmon	27	26	0.964285714
sugarpillpublishing	25	24	0.961538462
ellyblue	24	23	0.96
jessicafeinberg	24	23	0.96
239309591	22	21	0.956521739
gamesalute	67	64	0.955882353
eaglegryphon	20	19	0.952380952
maydaygames	39	37	0.95
674003445	32	30	0.939393939

Fig 2.17 Top 20 Kickstarter creators

The success rate of each creator is used as a very important feature for the following prediction models.

III. PREDICTIONS

3.1 Prediction on the success of a project

Given the data we have already crawled, we find that there are some useful data that we could use to derive some more predictions. First of all, we find that predicting the possibility of success of a project is a great idea. The goal is to figure out if a project will succeed or not, by using multiple factors.

Here are the list of factors that we think might affect the rate of success for a project:

Feature	Description
Number of Backers	More backers means more people get involved and such projects might have high potential to succeed.
Number of Comments	Similarly, more comments means there are more people interested in the project or actively engaged into the fundraising process, thus has the potential to get pledged more.
Number of Updates	If a project is working as scheduled, there should be constant updates. This feature indicates that the creator is actively making this project into happening.
Days that spend to pledge	If the duration of a project is set too short, there might not be enough time to fulfill the goal. There seems to be some relationships between duration and amount of goal.
Amount of Goal	As is concluded previously, the average goal is half the size of the average of pledged for successful projects. So the comparison of current pledged amount and goal, might give some hints about whether the project will succeed or not.
Amount of pledged	If current pledged amount is close to the goal, there is a high chance that a project will succeed. But it might be affected by how close the observed date is to the deadline.
Category Information	There are multiple categories and each category is inclined to have a different success rate. For example, it might be easier to achieve success by starting a gaming project than starting a craft project. The reason is various, maybe just because there are more people like playing games.

Geolocation Information	There are density difference among the projects across the globe. The location might affect the success rate as well. For example, a project in Los Angeles is likely to succeed since it has a high population and can easily attract more attention since it is a project from a bigger city with moure population. On the other hand, a project in Ohio may not get as much attention.
Length of Blurb	If the project is really prepared, and creators dedicate more into the project by illustrating more details on the project, it is more likely to get more attention from the public. Thus, we use this as a feature to indicate how dedicated the creators is to this project.
Profile of the Creator	The influence of the creator also account much. A successful creator who has previously got successful projects, are likely to get reputation and popularity among previous funders, and is more likely to accomplish new ones than a new creator since he might have more people know him (kind of like fans) and are willing to pledge for his/her projects. We use (1+success_num/ 1+total_num) to estimate the confidence of the possibility of the success rate of a creator. This is used to help estimate first seen creator with a success rate set to ½ .

3.1.1 Basic prediction

Based on the crawled data we already have, we decide to choose portion of features from the table above.

First we come up with feature sets: number of backers, number of comments, number of updates, duration of the project, goal and pledged amount of the project. These six features could describe the project in very detail.
The labeled data is split by the ratio of 7:3, where 70% is used as training data, and 30% is used to test the trained model. By using decision tree, classifier as well as logic regression classifier, we figure out there is a very high rate(98%) that the classifiers could correctly classify the testing dataset.

3.1.2 Importance of different features

Then we try to figure out the importance of each feature by eliminating number of features to choose.

Later we conclude that the high correctness ratio can be explained by the input parameter of goal and pledge. It is kind of obvious that by the time of deadline, if the pledge amount is larger than the goal amount, we can safely conclude that the project is successfully funded. Thus we need to extract different meaningful feature as the input parameter.

3.1.3 Prediction Update

As stated above, the basic prediction is more likely to be a summary rather than a prediction, thus we need some modification on the input features, and we need to figure out when the prediction takes place and when is the appropriate time to do the prediction. After consideration, we figure out that we need to do the prediction at just the beginning of this project. We made the decision based on the following reasons:
1. The average duration of a successful fund is around 32 days. Since the data is crawled each month, it is unlikely to figure out the trend for the procedure of a funding process.
2. It is more meaningful to predict at the starting stage of the project than the end of the project.
3. There are more features that we could use to predict, such as the category information, geolocation information, and creator profile information. The pledged amount, comments and update features describe the final state of a project, which is not appropriate to be used as prediction features. So we discard these two features.

As we conclude from the summary, some features have obvious correlation with the final state of a project:
1. The success rate for different categories varies a lot. A project related to filming is very likely to succeed, while a project on craft is much less likely to get enough fund. So we need to take this fact into consideration.
2. As for different places, such as in CA or NY, since they are more populated places, projects started there are more likely to get spread and get more attention from people thus are more likely to succeed, while cities in the central of the US, might be less likely to be fully funded. Thus, geolocation information also counts.
3. The creator himself may also have an impact on the success rate as well. A creator with lots of successful funds before is more likely get fully funded in the future as well since he have already earned reputation and trust across this environment, thus more people would pledge him again. We may not have information for a creator who just started his first funding project. We would like to use the success rate to indicate the ability of the creator. In order to eliminate 0 possibility on new creators, we use the following formula:

$$\frac{1+successfulProjects}{1+totalProjects}$$

With this formula, a new creator would get a possibility of ½ to achieve success in the next project. Overall, the more successful projects a creator has, the higher this feature value the creator would gain.

Since we are now doing predictions at the beginning of the project, some previous features used may not be available anymore. Now we use the following new features as the training data:
1. Length of the blurb
2. Successful rate on category
3. Successful rate on geolocation

4. Successful rate on the creator himself
5. Duration on the project
6. Amount of goal it is going to pledge.

The results are shown below:

Method	Correctness rate
Neuro Network	97.8%
Decision Tree	98%

The results shows that we could almost distinguish whether a project will succeed or not. Since we have divided dataset into training set (70%) and test set (30%), the precision of this model is reliable. In other words, the features we selected did a great job to illustrate the dataset.

3.1.3 Conclusion

As we can see, different aspects affects the possibility of whether a project will succeed or not. We don't have to do the prediction during the process. We achieve good precision on classifying at the beginning of a project.
So, as we have conclude from Section 2, people should take those factors into consideration before they really beginning to launch a project. For example, the ratio of success for his area (category), for the subcategory, as well as the creator's projects history. Also, think about the scale of the project (goal amount), as well as the duration (whether there are enough time to pledge) etc.
There are also other factors that should be considered into. For example, the backers' influence. Certain backers may very good at select projects that are more likely to succeed, and their choice of pledging of not might be a very good feature. However, the Kickstarters does not provide such data to the public, we are not able to analyze this feature during this project. Based on the crawled data, this is all we have, and it seems to predict well.

3.2 Prediction on the amount of money pledged eventually

In this part, we will describe our methods to predict the eventual amount of money that a project will be pledged. As we can imagine, this goal is quite ambiguous and very hard to predict, since there are a large variety of factors to take into consideration.

There are other factors that may have great impact on the result, for example, the popularity of the creator among his connection. It is quite normal that a famous people might get easily pledged 600% more of his original goal, while other new creators who started a new project, does not have a great chance to reach the amount of goal. It is quite hard to determine whether

a creator is popular or not. The success rate we got previously could not fully represent the creators' popularity. Apart from that, the various scales of projects also make prediction of pledged amount difficult.

However, we think it is worth giving a try. Based on the features we have summarized in Section 2, we try some popular models to make the predictions.

During this part, we use normalized training data. And basically we tried out two sets of training set, and the results are shown below.

3.2.1 training set 1

In this part, we try to do predictions at the beginning of the project, where there is obvious less clue about the condition of how the project goes. We choose the following columns: goal, creator_success_rate, subcategory_success_rate, us_state_success_rate, and period. These features are basically the same as the features that we used in the classifying model. And this is done by using both linear regression and neural network.

The following charts show the distribution of labeled data and predicted data. The factors on x axis are the normalized input data.

Comparing the following charts, we find that the similarity of distributions in the first chart is much higher than the others. There seems to be a strong connection between the creator and the pledged amount.
From the chart of periods, state and subcategory, we can find some sort of similarities, but not very obvious. The relation decremented when the final pledge value is large.
The goal, on the other hand, seems to have less effect on how much a project would be pledged.

Overall, the predicted amounts are not accurate, but still give us some hints about the correlations between features and pledged amount. We assume that the features we choose (mostly available at the beginning of the project) might not be good enough to do a precise prediction. Since some factors during the process of the pledging could affect the final pledged money. We did some changes on the next part.

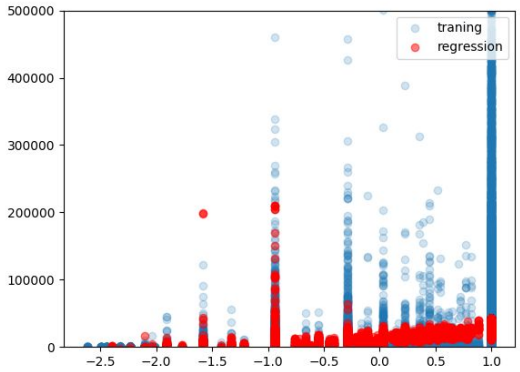

Fig 3.1 Creator vs pledged

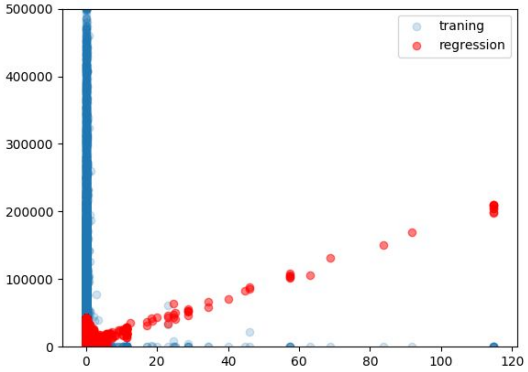

Fig 3.2 Goal vs pledged

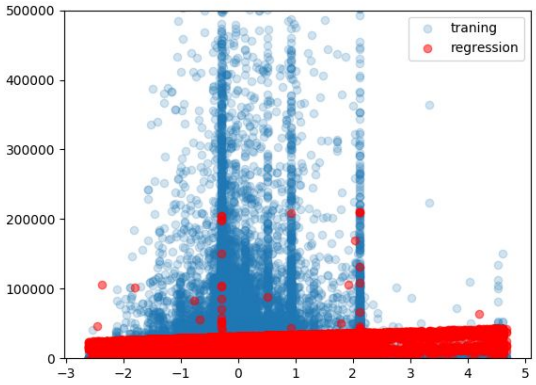

Fig 3.3 Period vs pledged

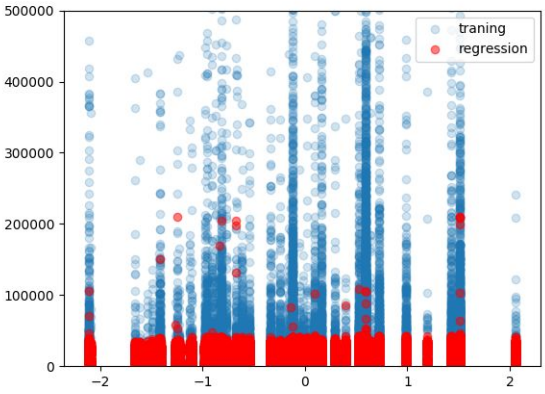

Fig 3.4 US state vs pledged

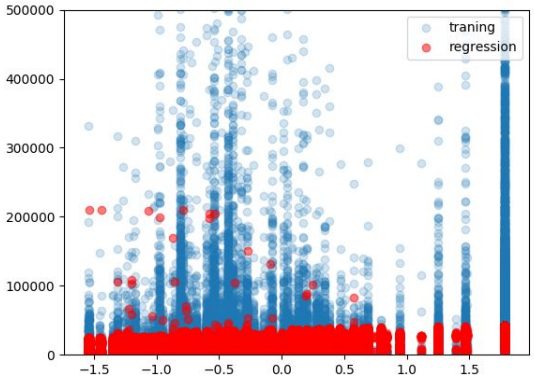

Fig 3.5 Subcategory vs pledged

3.2.2 training set 2

In this part, we add some features that are accessible during the pledge.

As we have calculated, the average pledge duration of a project is around 31 days. But data is crawled each month, thus it is unlikely to figure out the trend for each project. We try to use the data at the end of a project to find the factors that affects the final pledge amount.
We choose the following features: backers_count, comment_number, goal, period, update_number.

These features can briefly describe how many people are involved in the funding, how the creator actively engaged during the process of the funding and whether people are interested in this project or not.

After comparing the distributions of those charts, we find that amount of goal, number of updates and comment number show relatively strong connections with the pledged amount.
In the chart of backer number, it is obvious that the predicted set and training set don't fit well. We think it is probably because the pledged amount depends more on the quality rather than the quantity of backers. The amount of money one pledged may be different drastically.

To conclude, the number of backers may not necessarily be an important factor. Other factors, however, seems to be closely related to the final results. Since the average error of this model is much less than the previous one (17k :), we think the results in this part more convincing and worth considering.

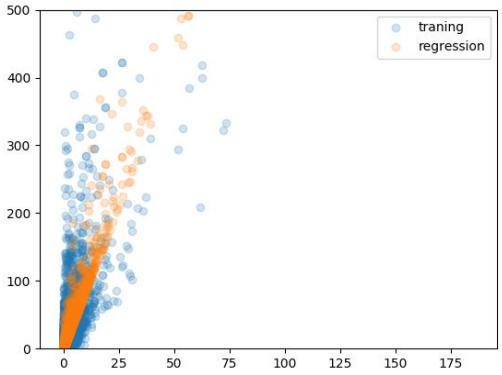

Fig 3.6 Backer number vs pledged

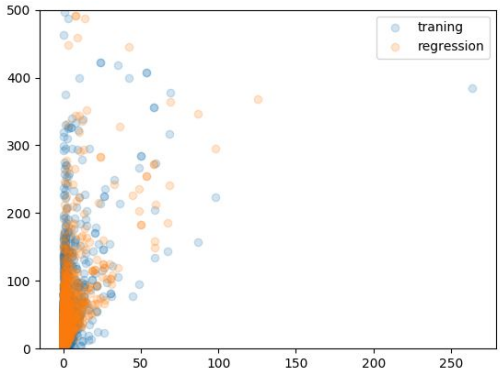

Fig 3.7 Comment number vs pledged

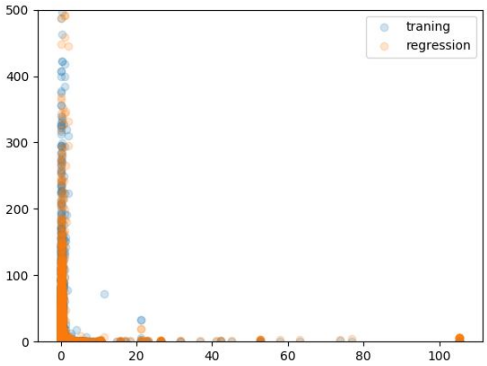

Fig 3.8 Amount of Goal vs pledged

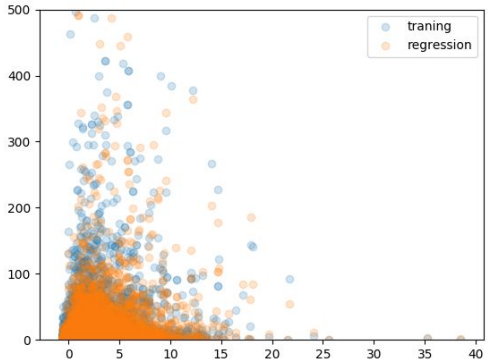

Fig 3.9 Duration vs pledged

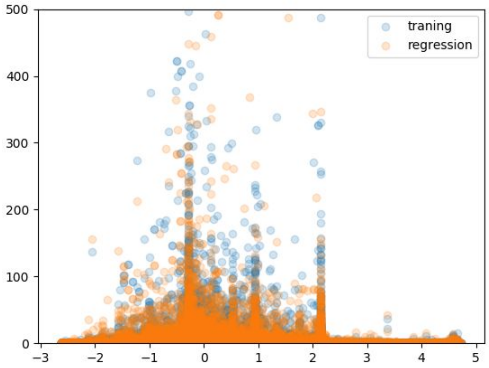

Fig 3.10 Number of updates vs pledged

3.2.3 Conclusion

Our classifier achieve great precision when used to predict whether a project will succeed or not. All features used in the model are available at the beginning of the project. Starters can use it as a reference.

The prediction of pledged money amount is not precise and it was expected before we did this experiment. However, it still provides some detailed measurements about the correlations between features and the pledged amount.

As compared with the first version, the precision of the second regression is much better, we conclude that number of comments, number of updates, and amount of goal have significant impact on the pledged amount excessively. However, these features is unlikely to obtained at the beginning of the project.

So, our advice for predicting the pledge money, is to pay attention of these features listed above during the process of a project.

For creators, in order to get a higher pledge, they need to actively engaged into the process (more updates), and interact more with people. Even with a small amount of goal, a project may get triple, or even 100 times pledged money more than the target goal. It is wiser to set a goal that reasonably lower than the expected goal.

IV. APPENDIX

4.1 Data source

We start from an open source dataset called ‘Web robots’. This website crawls kickstarters regularly and stores featured information of each project. You can find them here https://webrobots.io/kickstarter-datasets/. You can use shell script to batch download them.

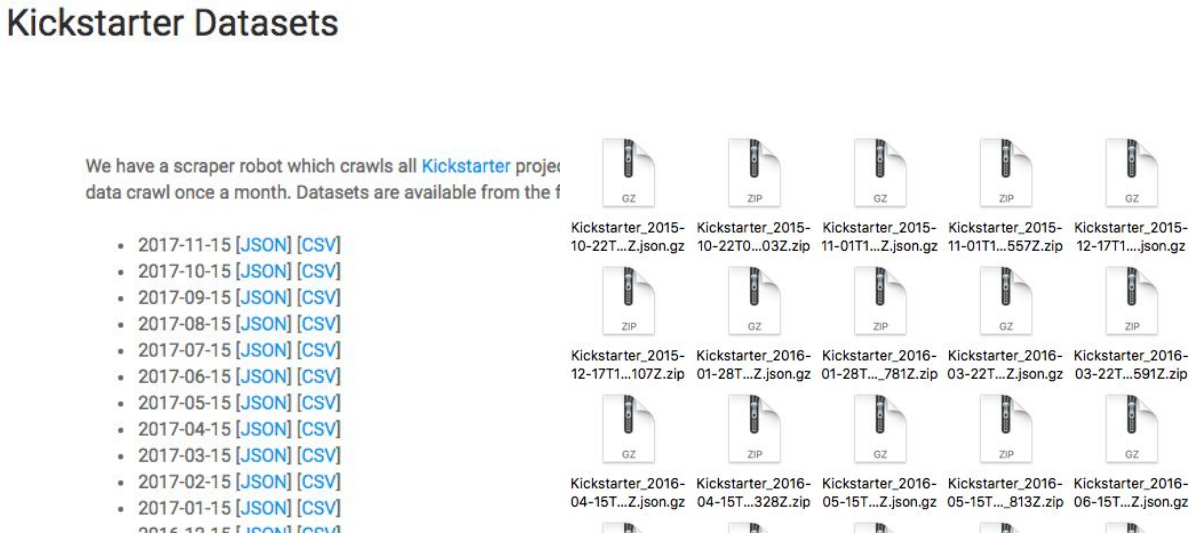

Fig 4.1 Dataset overview

The dataset contains many useful information, but our project needs some more specific features such as comments, comment number and update number. So we build our own crawler to retrieve those data.

4.2 Web crawler

4.2.1 Project urls

It is not an easy job to crawl all projects from the website starting from scratch. BFS is not efficient enough to do this job. Luckily, the dataset from ‘Web robots’ contains url information for each project. We use a python script to extract all urls from the dataset and build a url table on disk.

4.2.2 Crawling

We use a web crawler framework called ‘Scrapy’ for this project. Our crawler can crawl kickstarters using the url table we get before, retrieve information we need from web pages and store them in mongoDB with the url of each project as a key.
To speed up the crawling, our crawler runs in multi-threads and each thread would get a random IP and agent head from a pool. Also, they would change their IP and agent regularly to avoid being blocked by kickstarters.

4.2.3 Data storing

Data crawled using our crawler will be merged with dataset from ‘Web robots’ in mongoDB. MongoDB can export customized csv dataset as needed for summary and prediction usage.

4.3 Package usage overview

To get Web robots dataset and get all project urls, in dataprocessor:

```
sh start.sh
python3 project_url_util.py writemongo
python3 project_url_util.py writeurls
```

To run webcrawler, move urls.txt to kickstarterCrawler. Then, in kickstarterCrawler:

```
scrapy crawl kickstarter
```

All data are now in the mongoDB, we can get customized csv files we need using mongoDB export utils. MongoBooster is recommended to use.

4.4 Data summary

We did the summary based on the features that has previous discussed in Section 2.
The summary is done by using tools like: Spark, numpy, and pandas etc. We use Spark as the data aggregation tool to retrieve necessary data and use python to calculate (convert) the computed data to what we need in the prediction.

4.5 Predictions

There are mainly several approaches that we choose to do the prediction: Linear regression, Logistic Regression, Decision tree, Neural Network etc.
The reason to choose linear/ logistic regression is due to its easiness to implement
And we eventually choose Decision tree and neural network is based on the reason that these two could perform better at multiple input features. Decision tree is good at distinguish different features. Also, neural network is capable to changing parameters, such as changing architecture of the net in order to achieve a better performance. And it turns out they did pretty good job in classifying the successful projects.

Reference

Github link: https://github.com/liuchang0920/WSE-Project